Detect Profane Language in Streaming Services to Protect Young Audiences

Jingxiang Chen
Amazon Prime Video
jxchen@amazon.com

Kai Wei
Amazon Alexa
kaiwe@amazon.com

Xiang Hao
Amazon Prime Video
xianghao@amazon.com

Abstract

With the rapid growth of online video streaming, recent years have seen increasing concerns about profane language in their content. Detecting profane language in streaming services is challenging due to the long sentences appeared in a video. While recent research on handling long sentences has focused on developing deep learning modeling techniques, little work has focused on techniques on improving data pipelines. In this work, we develop a data collection pipeline to address long sequence of texts and integrate this pipeline with a multi-head self-attention model. With this pipeline, our experiments show the self-attention model offers 12.5% relative accuracy improvement over state-of-the-art distilBERT model on profane language detection while requiring only 3% of parameters. This research designs a better system for informing users of profane language in video streaming services.

1 Introduction

Streaming services such as Netflix and Prime Video have dramatically changed the media habits of young people, with six-in-ten primarily watching television today with streaming services (Pew, 2017). The increased exposure of online content has raised concerns about profane language appeared in these contents (Chen et al., 2012; Phan and Tan, 2017; Obadimu et al., 2019). Exposure to profane language can increase aggressive thoughts, angry feelings, physiological arousal, and aggressive behavior (Bushman, 2016; Phan and Tan, 2017) .

Profane language is a type of language that includes dirty words, swearing, and obscenity contents. Previous research has focused on developing automated techniques to detect profane language in user generated contents on social media. For example, there have been growing interests in detecting hate speech and racism on Twitter (Xiang et al., 2012; Badjatiya et al., 2017; Lozano et al., 2017). Some recent works have also studied offensive contents in Youtube (Alcântara et al., 2020). However, few studies have focused on profane language detection in streaming services that host movies and TV shows.

Recent works have shown the importance of data techniques such as pre-processing and augmentation in improving machine learning models. For example, there has been research on applying transfer learning or semi-supervised learning for learning word embedding and addressing insufficient data issues in tasks with limited sample sizes (Howard and Ruder, 2018; d'Sa et al., 2020). In addition, using text pre-processing methods such as text normalization, lowercasing, lemmatizing, tokenizing and multiword grouping can help increase sentiment, topic and polarity classification accuracy (Satapathy et al., 2017; Camacho-Collados and Pilehvar, 2017). However, few studies have focused on improving data techniques to better handle long sequence of text appeared in streaming videos. Research on addressing data issue have primarily focused on improving data quantity, rather than quality. Also, our problem has its novelty in that the data sets of most previous studies are on written text, which can have a different distribution from the spoken-form video captions.

In this work, we study the problem of profane language with a specific online video streaming service, Amazon Prime Video (PV), as an example. Specifically, we develop a data pipeline that can be integrated sentence level model to automatically predict the level of profanity in video titles according to their caption contents. We collect data from both the targeted service and public data set, and augment training data by merging multiple data sources. Our experiments show that this data collection pipeline that can be used to address long

Proceedings of the 4th Workshop on e-Commerce and NLP (ECNLP 4), pages 123–131

sequence of text and help non-hierarchical models to achieve state-of-the-art performance. Using this pipeline, we train a multi-head self-attention model on embedding pre-trained on PV caption dataset, and show this simple self-attention model (with 2 million parameters) can outperform the pre-trained distilBERT model (with 66 million parameters) that is fine-tuned on the same dataset by 9% accuracy.

2 Related Work

Profane language is a type of language that includes dirty words, swearing, and obscenity contents. Previous research in this area has primarily focused on detecting profane language in social media. For example, a recent work studied the diffusion of profanity in Sina Weibo, one of the largest Chinese social media platforms (Song et al., 2020). Research on abusive and hate speech detection (a close related research area to profane language detection) has focused on developing automatic techniques to identify racists and sexist on Twitter (Badjatiya et al., 2017; Lozano et al., 2017), Reddit (Chandrasekharan et al., 2017; Mohan et al., 2017), and Youtube (Obadimu et al., 2019). However, few studies have focused on detecting profane language in video stream services such as Netflix, Hulu, and Prime Video.

Research on this area has also shifted from using traditional machine learning methods to using deep learning methods. For example, while early work uses traditional machine learning classifiers, such as logistic regression, support vector machine, and tree-based methods (Xiang et al., 2012; Warner and Hirschberg, 2012), there has been a growing interest in applying LSTM, CNN, and BERT (Bidirectional Encoder Representations from Transformers) for detecting racism, sexism, hate, or offensive content (Badjatiya et al., 2017; Founta et al., 2019; Basile et al., 2019; Mozafari et al., 2019). However, one of the limits of these deep learning method is its capacity of working memory. This is because processing long texts (even for BERT) will quadratically increasing memory and time consumption and slicing the text by a sliding window or simplifying transformers, suffer from insufficient long-range attentions or need customized CUDA kernels (Ding et al., 2020). Due to this issue, these prior methods are not directly applicable for detecting profane language in video stream services since the video captions are often very long, with an average length of hundreds of sentences per video.

3 Data

Our data collection pipeline aims to address two challenges. First, there is no labeled data for profane language detection task in streaming service. To address this challenge, we collect data from a popular streaming service, design an annotation guideline, and hire human annotators to label the levels of profanity in video titles. Then, we augment training data by collecting additional labeled contents from a publicly available database. Last, we perform pre-processing on the collected data to improve its quality. Second, the title caption is so long that it is very challenging for a model to learn from context. To address this challenge, we develop a pipeline for creating sentence-level profanity labels using domain knowledge and title-level labeling information. We provide details about our data collection pipeline in the sections below.

3.1 Data Collection and Annotation

Figure 1 below shows the overall steps of our data collection and pre-processing technique pipeline. Specifically, we collected 150k titles that occupy the top streaming volume from Prime Video, a stream service that hosts movie and television shows. The titles include popular movies and TV shows in the most recent decades from global marketplaces, with duration ranging from 10 minutes to 3 hours (with 70% as TV). The titles are diverse and come from 200 genres such as kids cartoon, drama, romance, and horror. Their caption lengths range from a few sentences to thousands with average at about 700 sentences. We randomly sampled 5,260 titles from them for annotation.

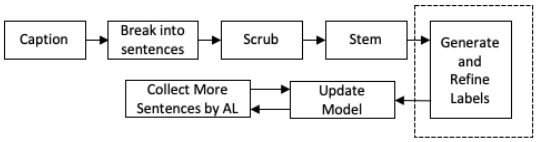

Figure 1: Data Collection and Pre-processing Pipeline

We develop codebook based on standard policy from movie and TV series rating associations, including Motion Picture Association of America (MPAA) and TVPG (TV Parental Guidelines Monitoring Board). In the codebook, we define the level of profanity in video titles based on their captions used. In total, the codebook includes 92 keywords and instructs on how the usage of them can lead to profanity. Some keywords always have mali-

cious meaning while others depend on their contexts. Based on the keyword frequency and the context severity, each title has a label from one of the following categories: None (all age), Mild (7+ kids), Moderate (13+ kids), Strong (16+ young adults), Severe (18+ adults), indicating the severity of profane language. For example, the singular instance use of disparaging slurs in the captions of a movie, such as Fag or Faggot in racism context, is rated as strong; and more than singular use of disparaging slurs is rated as severe.

A total of 15 human annotators are recruited to label the video captions. When annotating the captions of a video, the annotators go through the context of each keyword and decide whether the keyword is profane or not. After checking all such keywords in the entire caption, they make the final decision on the severity of the video title based on the counts and severity of the keywords appeared in the captions of a video. The maximum severity across all keywords in the captions of a video is used as the title-level rating. Throughout the whole annotation process, we randomly audit their labeling results to ensure the labeling quality.

3.2 Training Data Augmentation

Considering the small sample size collected in the above process, we add additional training data by collecting labeled data from the MPAA database according to the reason code description. We name the technique as *data augmentation* in this paper. For example, a video title may be rated as R due to strong language based on the reason code in MPAA database. In particular, we first select all the titles that is either G rating (suitable for all ages) or have profane language in the reason code. Then, we convert the rating from the MPAA standard to our rating category following a pre-defined mapping (i.e., G to None, PG to Mild, PG-13 to Moderate, NR/R/NC-17 to Severe). Note that MPAA does not have a rating corresponding to Strong (i.e., 16+ young adults) level. We clean the labels by comparing the rating with IMDB user votes when available, filter out mismatches. In total, we collected 5,010 additional labeled titles and combined them with the training data collected in Section 3.1.

3.3 Generating Sentence Labels

To generate sentence level labels, we use a keyword approach to scope relevant sentences in a video given its title level labels. Specifically, we use pre-defined keywords that express profanity to determine whether a given sentence is relevant to profanity. In total, there are 92 unique keywords. Some keywords always have malicious meaning while others depend on their contexts. For simplicity, we name the first class as *unambiguous* keywords and the second as *ambiguous* keywords. To identify them, we define the precision of a certain keyword w_i as the positive rate (i.e., the proportion of titles labeled as above all age) of training titles that contain w_i. We categorize a keyword to be unambiguous as long as its precision exceeds a given threshold ϵ ($\epsilon = 0.95$ for our case considering the chances that annotators may miss certain keywords and mislabel contain titles as all age). We ask domain experts to confirm the categorization results.

We then use a build-in sentence tokenizer from *python* package *NLTK* to break captions, which mainly relies on punctuation. We label the sentences that contain unambiguous keywords as positive (i.e., profane), and those do not as negative. This is because unambiguous keywords are profane in regardless of their context. In addition, we only cover the candidates from the pre-defined keywords, as recommended by domain experts.

The generated labels can have noise due to the unavoidable imperfection of the generation process. Inspired by knowledge distillation (Hinton et al., 2015), we train an intermediate classifier to generate the probability scores on sentences and save the predicted probability as a soft target to alleviate label noise. Such a new target contains the knowledge transferred from the intermediate classifier, and can be more robust than the original binary label. For example, it can help correct some labelling noises caused by limited rule coverage. As an example, the sentence *this guy sucks* is more likely to have a high score even if the previous rules do not cover it precisely than other similar expressions such as *he sucks* and *it sucks*. In this way, labelling it with a prediction score can be better than with zero. In practice, the intermediate classifier we use is a multi-head attention model, which will be introduced in Section 4.

3.4 Reduce Labeling Noise

We find that the model trained above still performs not very well on sentences with certain keywords, which may be caused by labeling noise. To address the issue, we apply the idea of active learning and manually label sentences picked by model predictions. First, we pick all the sentences that are

Severity	Count	Perc	# Sents	# Words
None	4061	55.9%	306	2634
Mild	971	13.4%	530	4812
Moderate	801	11.0%	720	5013
Strong	221	3.0%	720	5013
Severe	1216	16.7%	1071	8848

Table 1: Statistics of the training titles: number and percentage of titles at each rating, and the average number of sentences and words per each caption.

Severity	Count	Perc	# Sents	# Words
None	1036	35.4%	371	3161
Mild	398	13.6%	624	5301
Moderate	897	30.7%	703	5680
Strong	482	16.5%	851	6792
Severe	113	3.9%	1022	8772

Table 2: Statistics of the evaluation titles.

predicted as positive in none titles and those containing keywords from negatively predicted titles (i.e. all sentences of that title are predicted as negative). Second, we calculate the frequency n_i of each keywords from this sentence pool, and pick the top K ambiguous ones. Third, we randomly pick N sentences for manual labeling (N=2k for our case). In particular, we sample $N \cdot n_j / \sum_{i=1}^{K} n_i$ for each keyword j in the top K words obtained above. Finally, we label those N sentences, replicate them by T times to increase the weights (T=5 in practice), and combine them with the old training set for retraining. The newly added data helps correct the model at the boundary region.

3.5 Data Description

Table 1 and 2 present the overall statistics of training and evaluation data. Among these categories, None category has the shortest captions on average because many of the videos are kids cartoons or mini shows, which are often very short. In addition, the training and evaluation caption lengths are close to each other at each rating. However, there is a shift between the rating distributions of training and evaluation set. For example, the MPAA titles have more None labels than PV video dataset. The underlying reason can be: 1) MPAA may have less restrictive policy in labeling; and 2) PV video dataset may contain movies with larger diversity, and hence not dominated by None titles. The training set also has less Strong and more Severe titles because MPAA titles do not have Strong according to the rate mapping that we use.

4 Experiments

We integrated several models to our data pipeline and conducted experiments at both title level and sentence level. For the title level methods, we include xgboost and logistic regression, and deep learning methods such as an augmented version of the hierarchical attention network. We will introduce more details of these methods below. For the sentence level methods, we apply DistilBERT, rule based method that is used to create labels, and a sentence level multi-head attention model with and without the knowledge distillation soft target step. The purpose is to check whether the model learns the context information well, and whether the soft target helps. For each method, we have also fitted the model with only the 2.6k titles from prime video to study the effects of augmented data from MPAA.

4.1 Title Level Models

TF-IDF with traditional ML First, we use term frequency-inverse document frequency (TF-IDF, Leskovec et al. (2014)) to extract features and build models on them. We calculate the TF-IDF weights for unigrams and bigrams that have total frequencies greater than 5 and are contained by less than 90% of the titles (i.e., removing stop words). We try two classifiers logistic regression with L_2 penalty and xgboost (Chen and Guestrin, 2016), with unigram features alone (TI1 and TI3 in Table 3 and Table 4) or with both unigram and bigram features together (TI2 and TI4 in Table 3 and Table 4). For the logistic model, a multi-class cross-entropy loss is used for multi-category rating prediction.

Hierarchical Attention Network In addition, to capture the contextual information better, we propose an adjusted Hierarchical Attention Network (HAN, Yang et al. (2016)) on title level data. To enable HAN to take sentence level information, we propose generating 2k synthetic titles accordingly. Specifically, each title only contains the labelled sentence at a random position $\boldsymbol{s}_i$ and has other sentences as empty strings. Then, we fit the model with the synthetic titles and the old training set together. We include the 2k manually labelled sentences in the training set.

4.2 Sentence Level Models

DistilBERT We apply DistilBERT (Sanh et al., 2019) to the generated sentences, a distilled version of BERT that is 40% lighter but 60% faster and

can preserve over 95% of BERT's performances in downstreaming tasks. We fine tune the pretrianed DistillBERT on our sentence level data with generated labels. We do not apply token stemming because the model was pretrained on the raw corpus in a self-supervised fashion, using the BERT base model as a teacher.

Multi-Head Self-Attention Model We modify the self-attention model (Lin et al., 2017) to predict sentence-level profanity by introducing multi-head attention and soft target. The architecture of the model is presented in Figure 2. The model first converts each word ($\boldsymbol{w}_i$) of a n-length sentence into a vector with GloVE (Pennington et al., 2014) trained on the training captions of the 150k PV captions. Then the output is fed into a bidirectional GRU (BiGRU) layer with ReLU as the activation function. We find a BiGRU performs slightly better than a bidirectional LSTM in practice. This layer aims at capturing the long-term word dependency. The output of the BiGRU units, denoted as $\boldsymbol{h}_i$, is then passed into an m-head self-attention layer that allows attending different keywords of a sentence in a flexible way. The n-length weights of the j-th attention head are calculated via

$$\boldsymbol{a}_j = \text{softmax}(\tanh(\boldsymbol{\alpha}_j \cdot H^T + \boldsymbol{b}_j)),$$

for $j = 1, ..., m$, where $H = [\boldsymbol{h}_1, ..., \boldsymbol{h}_n]$ and ($\boldsymbol{\alpha}_j$, $\boldsymbol{b}_j$) are the coefficients. The elements of the vector (i.e. a_{1j}, ... , a_{nj}) represent how important each word is to determine the label of the sentence for the j-th attention head. The output of the attentions, $\boldsymbol{S}_j$ for $j = 1, ..., m$, is calculated by taking the weighted average as:

$$\boldsymbol{S}_j = \sum_{i=1}^{n} a_{ij} \cdot \boldsymbol{h}_i$$

The m outputs are concatenated with a fully connected with fully connected layer with a $sigmoid$ activation function built on the top of it. The loss function is the cross-entropy but with label using the soft target, denoted as q_i, as described in the second to the last step in generating sentence labels. In this way, the corresponding loss can be written as $L(\boldsymbol{p}, \boldsymbol{q}) = \sum_i q_i \log p_i$ where p_i is the output of the fully connected layer.

The predictions at sentence level are used to generate the title level labels. We calculate the frequencies of each keywords within the title by summing the scores of positive sentences that contain them. Then we accumulate the counts of keywords at each

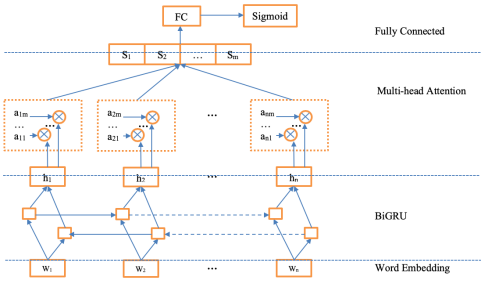

Figure 2: Sentence Level Multi-head Attention Model.

severity level following the standard policy that operators refer to. Such accumulated numbers are used to map the titles to the level of profanity. For example, over 10 times usage of non-aggressive coarse language and less than 2 times usage of disparaging slurs make the title labeled as strong.

4.3 Experiment Setup

For logistic regression, we choose the regularization parameter from $\{0, 1, 5, 10\}$. For xgboost, we tune the learning rate from $\{0.01, 0.05, 0.1\}$, max depth from $\{3, 6, 9\}$, number of estimators from $\{100, 200, 300\}$. For the multi-head self-attention model, we initialize the layer weights with Xavier uniform initializer and the bias with zeros. As to the hyper-parameter tuning, we use random search by selecting learning rate from $\{0.01, 0.001, 0.0001\}$ with decay rate of 0.9, RNN hidden size from $\{64, 128, 256\}$, attention output size from $\{64, 128, 256\}$, attention head number from $\{1, 3, 5\}$, and number of attention and dense layers from $\{1, 2\}$.

4.4 Evaluation Metric

We are interested in evaluating performance in terms of both binary classification (i.e. whether a video contains any profane language) and multi-class classification(i.e. levels of profane language in a video). For both, we use precision and recall as our primary metrics. In addition, it is important to control the chance of predicting a contain title as None because it is more risky to present an adult level video to kids than vice versa. In this way, we define a secondary metric in binary case as the recall at precision at 80%, 90%, and 97% for None titles, i.e. maximize the coverage of None titles at a given None precision level.

5 Results

In this section, we present our experiment results for detecting the presence of profane language and

the level of profane language for video titles.

5.1 Detecting Presence of Profane Language

Table 3 shows the model performance on detecting the presence of profane language in video title. Overall, models built based on the proposed pipeline of generating sentence labels (GSL) achieve better performance than those without it. Also, data augmentation helps in both methods with and without GSL (i.e., the best without DA always gets lower accuracy by over 3%). Among the method with GSL, multi-head attention model built on internally trained GloVE achieves the best accuracy and recall at precision of 97% and 80%. The comparison between intermediate and the final self-attention shows the marginal effect of the soft target. As to DistilBERT fine-tuned on PV data, it achieves the best recall at 90% precision but has the accuracy lower than the self-attention model trained on PV data by 3%. We also performed error analysis and found that DistilBERT prediction seems to do better at making inferences on semantic meaning, i.e., sentences that even do not contain keywords may still be predicted as positive if the expression is rude. However, DistilBERT does not perform well on ambiguous keywords; for example, it mis-classifies many negative sentences with *suck* as positive.

Compared to models built without generating sentence labels (w/o GSL), models built with GSL achieve better accuracy and recall at different precision threshold. This suggests that generating sentence labels can help improve model prediction on detecting the presence of profane language in video title. Within models built without GSL, the HAN model with data augmentation performs the best at recall at 97% precision, and the xgboost with both uni-gram and bi-gram features is the best among the traditional methods.

5.1.1 Error Analysis

The error analysis on the estimated word-level weights of HAN shows that title models assign high weights not only to the keywords related to profane language but also to those connected to other content descriptors like violence, such as *kill*, *police*, *liar* and *shoot*. The underlying reason can be that a severe title usually also contains elements such as violence and sexuality, and hence the existence of those corresponding words can be highly correlated. This can dilute the weights that are supposed to be given to the language keywords.

Method	Model	Acc (%)	R97	R90	R80
w/o GSL	TI1	78.3	36.1	54.5	67.6
	TI2	75.5	34.4	53.2	65.6
	TI3	85.2	53.1	78.1	96.8
	TI4	87.2	58.8	81.2	97.3
	HAN	85.5	58.9	74.6	89.1
	Best w/o DA	82.3	49.5	78.1	92.1
w/ GSL	Rule Based	87.7	76.3	-	-
	Intermediate	88.2	76.6	82.6	95.4
	DistilBERT	86.2	78.3	**86.4**	94.8
	Best w/o DA	87.4	76.9	81.7	88.4
	Self-Attn	**90.6**	**80.1**	84.1	**97.9**

Table 3: Overall Performance of detecting profane language: accuracy (Acc) and recall (Rec) at different precision thresholds (97%, 90%, and 80%). Among these methods, w/o GSL refers to the prediction method that use all captions to predict whether profane language presents in title-level without generating sentence level labels; and w/o GSL refers to the method with generating sentence level labels. TI1-TI4 represent the four baselines introduced in the experiment section. Best w/o DA means the best model that only uses 2.6k titles from targeted streaming services in training.

We also performed qualitative analysis to understand why the self-attention model built with GSL outperforms all the other models in both accuracy and recalls at all given precision levels. There can be two reasons. First, the soft target created by the intermediate model is more robust than the rule based target especially for the positives that go beyond the rule's coverage. Second, the manually labelled sentences picked from active learning results can help reinforce the weak signal for certain keywords. In particular, we find the model performs significantly better than others in the top five frequent words (*hel, jerk, suck, piss, ho*) picked by active learning. In addition, it is not surprising that the intermediate model outperforms the rule based slightly. The main difference comes from sentences with ambiguous keywords. The model can correct certain labelling issues by applying the average scores calculated by its fitted weights on these keywords and their neighbors. The best sentence model without data augmentation performs a bit worse than the rule based, which can be caused by smaller coverage of certain keywords in training.

In addition, the attention layer localizes the keywords well by assigning them with larger weights. In Figure 3, we pick both positive and negative sentences with certain ambiguous keywords when one-head attention is used. We print their scores as well as the attention weights (scaled to 10) on each word. The model learns the context in the Bi-GRU

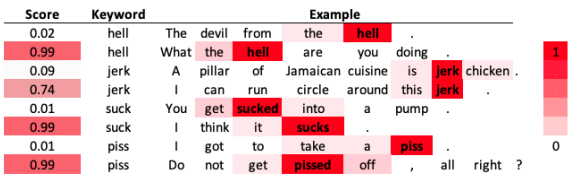

Figure 3: Sentence classification examples: confidence score and weight coefficients (scale to 10) at the attention layer at one head. According to the label, sentence 2, 4, 6 and 8 contain profane language.

Method	Model	Acc (%)	Prec (%)	Rec (%)
w/o	TI1	31.2	42.1	78.8
	TI2	35.6	45	79.1
	TI3	54.7	75.5	92.5
	TI4	53.4	73.2	89.8
	HAN	55.8	73.7	**92.9**
	Best w/o DA	38.2	64.4	81.8
w/ GSL	Rule Based	76.3	89.8	78.4
	DistilBERT	71.2	86.3	79.4
	Best w/o DA	72.2	87.7	79.8
	Self-Attn	**80.1**	**93.6**	80.2

Table 4: Overall performance of detecting levels of profane language: accuracy, precision and recall.

layer by passing the neighboring information to the keyword location, and thus equip the same keyword with different output vectors. For example, our observation shows the output vector $\boldsymbol{h}$ (Figure 2) for word *hell* are quite different between *The devil from the hell* and *What the hell are you doing*. In this way, the fully connected layer learns such difference made by context and predicts different scores.

5.2 Detecting the Level of Profane Language

Table 4 shows the overall model performance on detecting the level of profanity. The self-attention model achieves the best accuracy and precision, and it beats the rule based by 3% in accuracy, mainly due to a higher precision at Mild and Moderate level (the rule misses a certain amount of those titles due to the lack of coverage). Also, we find DistilBERT performs a bit worse than the rule based approach due to its suboptimal performance on ambiguous keywords. As to the methods without GSL, HAN achieves the best performance and recall even higher than the self-attention model built with GSL. This shows that it can be beneficial to build hierarchical models to predict title-level profane language from its long video captions when the proposed pipeline is not available. The reason why the HAN has a better None recall at the top severe level is that it tends to be more conservative in predicting a higher rate.

Label \Pred	None	Mild	Moderate	Strong	Severe
None	829	**116**	**78**	9	4
Mild	14	197	**179**	8	0
Moderate	3	16	844	32	2
Strong	2	0	40	433	7
Severe	1	0	1	37	74

Table 5: Confusion matrix of multi-category rating prediction.

We also find that none of the traditional models built without GSL performs well. In particular, they tend to overestimate the proportion of the severe level and underestimate that of the Moderate and Strong levels, possibly misled by the distribution shift between training and evaluation. In addition, the title level methods trained without data augmentation give significantly worse performance, indicating sample size is crucial when predicting the level of profane language in video title.

5.2.1 Error Analysis

The confusion matrix of our model is reported in Table 5. The main errors come from overrating some None and Mild titles, as the bold numbers show. Error analysis finds that some errors at Mild vs Moderate are caused by the shift of keyword frequency distribution from training to evaluation. For example, the words *crap* and *blow* are more often in evaluation set when used as negative words. In addition, the difference between Mild and Moderate can be quite subtle even to human operators. A deeper analysis shows a certain amount of errors on None titles may be caused by possible annotation mistakes, especially on the boundary of Mild and Moderate. A better way to evaluate the model performance can be using the distance between prediction and label to measure the loss (e.g., misclassifying Moderate as None should be worse than as Mild).

6 Conclusion

In this paper, we presented a data collection pipeline to generate the high quality sentence-level label for profane language detection in a streaming service. This pipeline included specific knowledge distillation and active learning ideas to refine such labels. We applied data augmentation, collected training data from both the targeted streaming service and public open source, and applied a workflow to fill the gap caused by the rating policy inconsistency. We built a multi-head self-attention

model for sentence level detection and aggregated the detections to title level for rating prediction. The experiment showed the proposed model outperformed all the baselines including the hierarchical attention network and DistilBERT, and also beat the rules that created the labels. In addition, the output attention weights showed success in locating the right keywords. Future research directions include the exploration on how the proposed pipeline will help detect more general profanity defined in multimodality, such as visual and audio.

Acknowledgements

We thank the anonymous reviewers for their feedback and insights in improving the work.

References

Cleber Alcântara, Viviane Moreira, and Diego Feijo. 2020. Offensive video detection: Dataset and baseline results. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4309–4319.

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

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Brad J Bushman. 2016. Violent media and hostile appraisals: A meta-analytic review. *Aggressive behavior*, 42(6):605–613.

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2017. On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. *arXiv preprint arXiv:1707.01780*.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–22.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80. IEEE.

Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Cogltx: Applying bert to long texts. *Advances in Neural Information Processing Systems*, 33.

Ashwin Geet d'Sa, Irina Illina, Dominique Fohr, Dietrich Klakow, and Dana Ruiter. 2020. Label propagation-based semi-supervised learning for hate speech classification. In *Insights from Negative Results Workshop, EMNLP 2020*.

Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM Conference on Web Science*, pages 105–114.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2014. *Mining of Massive Datasets*, 2nd edition. Cambridge University Press, USA.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Estefanía Lozano, Jorge Cedeño, Galo Castillo, Fabricio Layedra, Henry Lasso, and Carmen Vaca. 2017. Requiem for online harassers: Identifying racism from political tweets. In *2017 Fourth International Conference on eDemocracy & eGovernment (ICEDEG)*, pages 154–160. IEEE.

Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. 2017. The impact of toxic language on the health of reddit communities. In *Canadian Conference on Artificial Intelligence*, pages 51–56. Springer.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer.

Adewale Obadimu, Esther Mead, Muhammad Nihal Hussain, and Nitin Agarwal. 2019. Identifying toxicity within youtube video comment. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior*

Representation in Modeling and Simulation, pages 214–223. Springer.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pew. 2017. About 6 in 10 young adults in u.s. primarily use online streaming to watch tv. *Pew Research Center*.

Quang Anh Phan and Vanessa Tan. 2017. Play with bad words: A content analysis of profanity in video games. *Acta Ludica-International Journal of Game Studies*, 1(1):7–30.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ranjan Satapathy, Claudia Guerreiro, Iti Chaturvedi, and Erik Cambria. 2017. Phonetic-based microtext normalization for twitter sentiment analysis. In *2017 IEEE international conference on data mining workshops (ICDMW)*, pages 407–413. IEEE.

Yunya Song, K Hazel Kwon, Jianliang Xu, Xin Huang, and Shiying Li. 2020. Curbing profanity online: A network-based diffusion analysis of profane speech on chinese social media. *New Media & Society*, page 1461444820905068.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26. Association for Computational Linguistics.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.