A Hybrid Modeling Approach for an Automated Lyrics-Rating System for Adolescents

Jayong Kim[1] and Mun Y. Yi[2*]

[12] Graduate School of Knowledge Service Engineering,
Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
{kjyong}{munyi}@kaist.ac.kr

Abstract. The South Korean government operates human-based lyrics-rating systems to reduce adolescents' exposure to inappropriate songs. In this study, we developed lyrics classification models for an automated lyrics-rating system for adolescents. There are two kinds of inappropriate lyrics for adolescents: (1) lyrics with inappropriate words and (2) lyrics with inappropriate content based on the semantic context. To tackle the first issue, we propose $\log CD_{\alpha}$ as a method for generating a lexicon of inappropriate words. It attained the highest performance among the lexicon-based filtering methods examined. Further, to deal with the second issue, we propose a hybrid classification model that combines $\log CD_{\alpha}$ with an RNN based model. The hybrid model composed of a 'lexicon-checking model' and a 'context-checking model' achieved the highest performance among all of the models examined, highlighting the effectiveness of combining the models to specifically target each of the two types of inappropriate lyrics.

Keywords: Lyrics Classification, Offensive Language Detection, RNN

1 Introduction

To reduce the exposure of adolescents to lyrics that contain depictions of profanity, violence, sex, and/or substance abuse, the South Korean government operates human-based lyrics-rating systems. The Ministry of Gender Equality and Family (MOGEF) classifies lyrics as either clean or inappropriate for adolescents, and they are prohibited from accessing music records with lyrics that are considered as inappropriate.

The human-based lyrics-rating system includes the basic work of the monitoring staff, an initial review of the committee every other week, and a main review of the committee once a month [6]. Because the lyrics-rating system is a post deliberation, adolescents can still be exposed to inappropriate lyrics, particularly if the deliberation process takes time. Furthermore, a number of experts and resources are required continuously for the operation of the current lyrics-rating systems.

Automation of this rating process could be a valuable solution by saving time and resources. The purpose of this study is to develop an effective lyrics classification model for an automated lyrics-rating system. We believe that the developed models and the approaches could be easily applicable to the lyrics-rating systems of other countries.

2 Related Work

Chin et al. [3] studied an inappropriate lyrics classification model for the first time. They reported that there were two main types of inappropriate lyrics for adolescents: (1) 'lyrics that contain inappropriate words for adolescents' (Type I) and (2) 'lyrics that do not contain inappropriate words, but contain explicit content based on the context' (Type II). Although they noticed that there were two types of inappropriate lyrics, the authors did not develop a specific model that particularly dealt with them. In the present study, we developed a model that focuses on these two types of inappropriate lyrics.

It is easier to classify Type I lyrics than Type II lyrics because we only need to check for the presence of inappropriate words. The basic approach for tackling this issue is lexicon-based filtering (keyword matching) [16,18]. However, in the absence of well-defined lexicon data, it is difficult to apply this approach. Hence, automatic generation of a lexicon of inappropriate words for adolescents is a viable, practical solution. The approach has only been studied in a limited scope for social media content [1,12]. In the present study, we expanded it to lyrics and examined its applicability for the classification of Type I lyrics.

In order to classify Type II lyrics as inappropriate, the semantic context of the words needs to be grasped. A lexicon-based filtering approach cannot capture the context of words because it only checks for their presence in the lexicon and does not consider the other words. To understand the semantic context of words, recurrent neural network (RNN)- and convoluted neural network (CNN)-based sequential data processing models have been studied [7, 19]. Lyrics that contain even one single profanity can be classified as inappropriate according to the lyrics-rating system of MOGEF. These kinds of lyrics are Type I lyrics and RNN- or CNN-based model might not be suitable for them. These types of models can lose information on the existence of inappropriate words because they make low dimensional vectors of lyrics, not bag-of-words vectors. Therefore, to compensate for this weakness, we propose a hybrid classification model, which is composed of a 'lexicon-checking model' and a 'context-checking model' for classifying Type I and Type II inappropriate lyrics, respectively.

3 The Proposed Model

3.1 For Type I Lyrics

Automatic Lexicon Generation. Type I lyrics contain inappropriate words for adolescents, for example, *'I don't give a f***'*. This can be verified by checking whether the lyrics contain particular words in the lexicon of inappropriate words for adolescents. Filtering methods for feature selection, such as log odds ratio (LOR), correlation coefficient (CC), and supervised word weighting schemes such as relevant frequency (RF) [8] and LogCD [4] can be used for the automatic generation of a lexicon of inappropriate words. According to [15], the score that these methods give to a term $\boldsymbol{t_k}$ can be represented by the number of positive-class-documents with $\boldsymbol{t_k}$ and the number of negative-class-documents with $\boldsymbol{t_k}$.

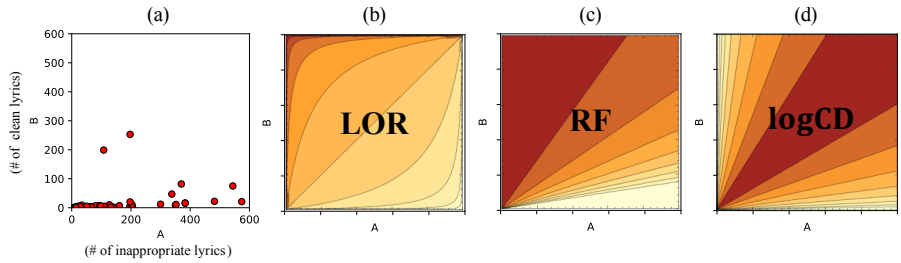

Fig. 1. (a) A scatter plot of 539 profanity words collected from [13, 14]. A axis is the number of inappropriate lyrics with a profanity word $\boldsymbol{w_k}$, B axis is the number of clean lyrics with a profanity word $\boldsymbol{w_k}$. (b)-(d) The scoring tendencies of the various methods according to A and B. The brighter the color in the contour plot, the higher the score.

In Fig. 1 (a), the profanities in the lyrics show a clear pattern in that most of the points appear near the A axis because lyrics with even a single inappropriate word should be classified as inappropriate. Therefore, methods that give a high score to the words near the A axis are suitable for the automatic generation of a lexicon of inappropriate words for adolescents.

$$\text{logCD}_\alpha = \log\left(\frac{\frac{A+\alpha}{\#\ of\ inappropriate\ lyrics}}{\frac{B+\alpha}{\#\ of\ clean\ lyrics}}\right) \tag{1}$$

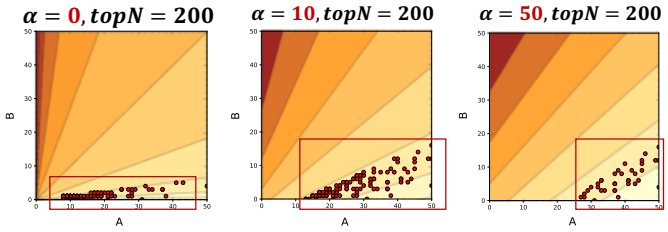

Fig. 2. Selected inappropriate words for adolescents with various α of $\textbf{logCD}_\alpha$.

We modified logCD [4] to generate a more effective lexicon (see logCD_α (1)). Absolute values in logCD were removed to make it assign a high score to words near the A axis but not the B axis. Further, we added α to the numerator and denominator. Fig. 2 shows that as α of logCD_α increases, words with a low document frequency are gradually excluded from the lexicon even though they are near to the A axis. On the other hand, as α increases, words with a high document frequency are included in the lexicon of inappropriate words, even when they appeared among the clean lyrics. As words appear more frequently, the more likely they are to appear in the clean lyrics. logCD_α considers this pattern and can give tolerance to the words by increasing α.

The Lexicon-Checking Vector. We can create various lexicons by changing the α of logCD_α and the number of words included in the lexicon. To make lexicon-checking

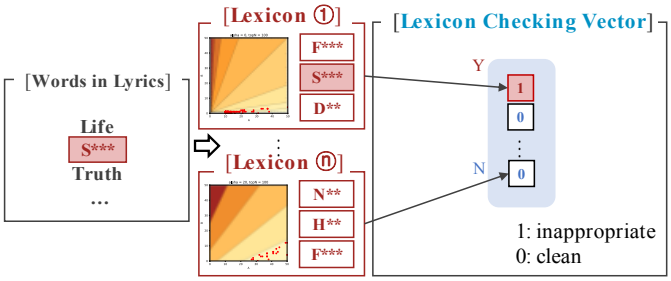

Fig. 3. The process of generating the lexicon-checking vector

vectors, lexicon-based filtering is carried out for each lexicon (see Fig. 3). If lyrics contain words in any one of the lexicons, the lyrics are classified as inappropriate. After conducting this process for all lexicons, predictions based on each lexicon are carried out to determine whether the lyrics are appropriate or not. We collected the top k predictions of the lexicons which performed the best when validating the data. This vector played the role of a 'lexicon-checking vector' in the hybrid model.

3.2 For Type II Lyrics

The Context-Checking Vector. Type II lyrics, for example, contain sentences like *'Take my skin off, cut out my belly'*. When we look at the overall expression, it might be inappropriate for adolescents. However, if we break the expression up into words like 'cut', 'out', or 'belly', it seems like these words are appropriate. That is to say, the semantic context of the words is important in Type II lyrics rather than the literal meaning of the words themselves. Therefore, we directly applied Hierarchical Attention Networks (HAN), which is an RNN-based model for sequential and hierarchical processing of words [19]. After training the HAN, the output value of the last layer before the Softmax layer was used as the 'context-checking vector' of the lyrics.

3.3 The Hybrid Approach for Type I and Type II Lyrics

To consider both Type I and Type II lyrics, we designed a hybrid classification model of inappropriate lyrics. The 'lexicon-checking vector' and the 'context-checking vector' of each lyric were concatenated into a single vector, after which a classifier learned these vectors.

4 Experiments and Results

The lyrics-rating results of MOGEF during 2010.1-2017.8 were collected from the MOGEF website [11]. Inappropriate lyrics from 7,468 songs and clean lyrics from 62,609 songs that did not have an 'Adults Only' tag on the music streaming sites were crawled from various lyrics databases. The class imbalance of the dataset was intended

to reflect a real-world lyrics-rating system. The dataset was split into training data, validation data, and test data with a ratio of 8:1:1. All of the hyper-parameters were tuned using the validation data. α was tuned between 0 and 100 and the number of words in a lexicon (topN) was varied between 10 and 400. The lyrics were tokenized using the Part-Of-Speech taggers from the NLTK and Komoran packages. Because of the class imbalance, the F1 score and area under the precision-recall curve (PR AUC) were used as performance measures [5,17].

4.1 Automatic Lexicon Generation

Table 1. The best results of the lexicon-based filtering of each method varying topN.

Method	topN	F1
$\text{logCD}_{\alpha=20,k=1}$	25	0.7562
log odds ratio (LOR)	300	0.7368
$\text{logCD}_{without\_absolute}$	400	0.6779
relevant frequency (RF) [8]	400	0.6779
Mubarak et al. [12]	21133	0.5330
correlation coefficient (CC)	10	0.5251
man-made dictionary [13,14]	539	0.4898

We conducted lexicon-based filtering for each lexicon generated by the diverse methods. If the lyrics contained any single token in a generated lexicon, it was classified as inappropriate. Table 1 reports that logCD_{α} attained the highest performance for lexicon-based filtering as it used only 25 words. This is 0.1% of the total number of words. In addition, logCD_{α} used fewer words than the existing methods that showed comparable performance. It suggests that logCD_{α} is efficient for generating an effective lexicon of inappropriate words for adolescents.

4.2 Hybrid Classification Model

We compared the proposed model with the bag-of-words model (TF-IDF), the document embedding model (Doc2Vec [9]), the topic modeling model (LDA [2]), and the lyrics classification model proposed by Chin et al. [3]. We tested various classifiers, namely AdaBoost, Bagging, and k-nearest neighbor (KNN), and we reported the KNN here because all of them showed similar results.

HAN produced the highest performance among the non-hybrid models (Table 2). However, the performance difference between $\text{logCD}_{\alpha=20,k=1}$ and HAN was 1%. Checking the presence of 25 words, logCD_{α} achieved a comparable performance at a lower cost when compared to the deep learning model.

The hybrid classification model based on HAN and logCD_{α} showed the highest performance among all of the models compared (Table 2). It outperformed its sub-models: logCD_{α} and HAN. The hybrid model showed a higher performance when the size of the 'lexicon-checking vector' was 100 ($\text{logCD}_{\alpha=20,\text{k}=100}$) compared to when it was 1 ($\text{logCD}_{\alpha=20,\text{k}=1}$), meaning that using various lexicons could improve its performance.

Table 2. The experimental results of the compared models

Model	F1	PR AUC
Doc2Vec [9] +KNN	0.5066	0.5986
TF-IDF+KNN	0.5299	0.5481
LDA [2] +KNN	0.6507	0.6720
Chin et al. [3]	0.7478	0.7774
HAN [19]	0.7665	0.8249
Hybrid(Doc2Vec+HAN)+KNN	0.7744	0.8117
Hybrid($\log CD_{\alpha=20,k=1}$+HAN)+KNN	0.7809	0.8275
Hybrid($\log CD_{\alpha=20,k=100}$+HAN)+KNN	**0.8049**	**0.8600**

The improvement achieved by the hybrid model might have been simply due to combining the multiple models. To check for this, we made many hybrid models with various combinations of single models. However, even the best other combination, Hybrid(HAN +Doc2Vec), showed little improvement or less performance than its sub-models, which indicates that the proposed combination of models specifically targeting Type I and Type II lyrics were synergistic.

5 CONCLUSION

Automating a lyrics-rating system can save time and resource relative to the current human-based system. In this research, we developed a hybrid model of lyrics classification for an automated lyrics-rating system. Extending the extant research, we focused on two types of inappropriate lyrics for adolescents.

To classify Type I lyrics, we first found the pattern of inappropriate words for adolescents. From the pattern, we developed insight into what kinds of scoring methods might be suitable for finding inappropriate words. This approach was then applied to other text classification areas where the class of the text depended on the presence of specific words in the content, such as profanity filtering.

We proposed $\log CD_{\alpha}$ as an automatic generation method for lexicons of inappropriate words, which can be further used to generate domain-specific lexicons regardless of language. In addition, $\log CD_{\alpha}$ showed the highest performance with the fewest words for lexicon-based filtering compared to existing methods, showing that it can be applied to areas where both time and resource savings are important, such as real-time inappropriate content detection with a large amount of data.

We designed a hybrid classification model that considers both Type I and Type II lyrics by learning the 'lexicon-checking vector' and the 'context-checking vector'. This hybrid model showed the highest performance among all of the models we examined. As the hybrid modeling approach considers both the lexicon and the context together, its performance could be assessed in other document classification tasks in which both checking the lexicon and determining the context are required.

References

1. Abozinadah, Jones Jr.: A Statistical Learning Approach to Detect Abusive Twitter Accounts. In: Proceedings of the International Conference on Compute and Data Analysis. pp. 6-13. ACM, New York (2017).
2. Blei, Ng, Jordan: Latent dirichlet allocation. Journal of Machine Learning research. 3(Jan), 993-1022 (2003).
3. Chin, Kim, Kim, Shin, Yi: Explicit Content Detection in Music Lyrics Using Machine Learning. In: Big Data and Smart Computing (BigComp), 2018 IEEE International Conference on. pp. 517-521. IEEE (2018).
4. Fattah, Abdel: New term weighting schemes with combination of multiple classifiers for sentiment analysis. Neurocomputing. 167, 434-442 (2015).
5. He, Haibo, Garci: Learning from imbalanced data. IEEE Transactions on knowledge and data engineering. 21(9), 1263-1284 (2009).
6. Hwang, Choi, Yoon: Study on how to improve operating system of the commission on adolescents Protection (청소년보호위원회 운영체계 발전방안 연구). http://www.prism.go.kr/homepage/origin/retrieveOriginDetail.do;jsessionid=A81566807529C919EE3FAD27DCD2DD56.node02?cond_research_name=&cond_research_start_date=&cond_research_end_date=&cond_organ_id=1382000&research_id=1382000-201300016&pageIndex=20&leftMenuLevel=120, last accessed 2017/09/01.
7. Kim Yoon: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. pp. 1746–1751. ACL (2014).
8. Lan, Tan, Su, Lu: Supervised and traditional term weighting methods for automatic text categorization. IEEE transactions on pattern analysis and machine intelligence. (31)4, 721-735 (2009).
9. Le, Mikolov: Distributed representations of sentences and documents. In International Conference on Machine Learning. Jan,1188-1196 (2014).
10. Mikolov, Sutskever, Chen, Corrado, Dean: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. pp. 3111-3119. Curran Associates Inc. (2013).
11. Ministry of Gender Equality and Family: In-appropriate media for Juvenile (청소년유해매체물). http://www.mogef.go.kr/sp/yth/sp_yth_f013.do, last accessed 2017/09/01.
12. Mubarak, Darwish, Magdy: Abusive language detection on Arabic social media. In: Proceedings of the First Workshop on Abusive Language Online. pp. 52-56. ACL (2017).
13. NamuWiki: Profanity/Korean. https://namu.wiki/w/%EC%9A%95%EC%84%A4/%ED%95%9C%EA%B5%AD%EC%96%B4, last accessed 2017-09-01.
14. NoSwearing: List of Swear Words, Bad Words, & Curse Words. https://www.noswearing.com/dictionary. last accessed 2017-09-01.
15. Ren, Fuji, Sohrab: Class-indexing-based term weighting for automatic text classification. Information Sciences. 236, 109-125 (2013).
16. Sood, Owsley, Antin, Churchill: Using Crowdsourcing to Improve Profanity Detection. AAAI Spring Symposium: Wisdom of the Crowd. vol 12, 6 (2012).
17. Sun, Yanmin, Wong, Kamel: Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence. 23(4), 687-719 (2009).

18. Xiang, Fan, Wang, Hong, Rose: 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. pp.1980-1984. ACM (2012).
19. Yang, Yang, Dyer, He, Smola, Hovy: 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. pp. 1480-1489. NAACL (2016).