Automatic Identification of Age-Appropriate Ratings of Song Lyrics

Anggi Maulidyani and Ruli Manurung
Faculty of Computer Science, Universitas Indonesia
Depok 16424, West Java, Indonesia
anggi.maulidyani@ui.ac.id, maruli@cs.ui.ac.id

Abstract

This paper presents a novel task, namely the automatic identification of age-appropriate ratings of a musical track, or album, based on its lyrics. Details are provided regarding the construction of a dataset of lyrics from 12,242 tracks across 1,798 albums along with age-appropriate ratings obtained from various web resources, along with results from various text classification experiments. The best accuracy of 71.02% for classifying albums by age groups is achieved by combining vector space model and psycholinguistic features.

1 Introduction

Media age-appropriateness can be defined as the suitability of the consumption of a media item, e.g. a song, book, film, videogame, etc., by a child of a given age based on norms that are generally agreed upon within a society. Such norms may include behavioral, sociological, psychological, and other factors. Whilst we acknowledge that this is largely a subjective judgment, and that there may be wide variance between very small circles that could be considered demographically homogenous, nevertheless, parents, educators, and policymakers may find such judgments valuable in the process of guiding and supervising the media consumption of children.

This topic is closely related to well-known content rating schemes such as the MPAA film rating system[1], but whereas such schemes are focused more on whether a film contains adult material or not, age-appropriatness can be thought of as being more nuanced, and takes into consideration more factors such as educational value.

One popular resource for such ratings is Common Sense Media[2], a website that provides reviews for various media, with a focus on age appropriateness and learning potential for children.

Whilst acknowledging that such ratings are of interest to many people, the position of this research is neutral towards the efficacy and utility of such ratings: we only seek to ask the question of whether it is possible to automate the identification of these age-appropriateness ratings.

This work focuses on song lyrics. There are many aspects that can contribute to the age-appropriateness of a song, but we believe that by far the most dominant factor is its lyrics. Thus, the approach that is taken to automating the identification of age-appropriatness ratings is to treat it as a supervised text classification task: first, a corpus of song lyrics along with age-appropriateness ratings is constructed, and subsequently this corpus is used to train a model based on various textual features.

To give the reader an idea of this task, Figures 1 to 3 show a sampler of snippets of lyrics[3] from songs along with their age-appropriate ratings according to Common Sense Media. Our goal is to be able to automatically predict the age-appropriate rating given the lyrics of a song in such cases.

Oh, I'm Sammy the snake
And I look like the letter "S"ssss.
Oh, yes.
I'm all wiggly and curvy,
And I look like the letter "S"ssss.
I confess.
(age-appropriate rating: 2)

Figure 1: Snippet of "Sammy the Snake", from Sesame Street Halloween Collection

[1] http://www.mpaa.org/film-ratings

[2] http://www.commonsensemedia.org

[3] All works are copyrighted to their respective owners.

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 583–587,
Beijing, China, July 26-31, 2015.

Do you want to build a snowman?
Come on, let's go and play
I never see you anymore
Come out the door
It's like you've gone away
(age-appropriate rating: 5)

Figure 2: Snippet of "Do you want to build a snowman?", from Frozen Original Motion Picture Soundtrack

You can take everything I have
You can break everything I am
Like I'm made of glass
Like I'm made of paper
Go on and try to tear me down
I will be rising from the ground
Like a skyscraper
Like a skyscraper
(age-appropriate rating: 9)

Figure 3: Snippet of "Skyscraper", from Unbroken - Demi Lovato

In Section 2 we discuss related work, before presenting our work on constructing the corpus (Section 3) and carrying out text classification experiments (Section 4). Finally, we present a tentative summary in Section 5.

2 Related Work

To our knowledge, there is no previous work that has attempted what is described in this paper. There is some thematically related work, such as automatic filtering of pornographic content (Polpinij et al., 2006; Sood et al., 2012; Xiang et al., 2012; Su et al., 2004), but we believe the nature of the task is significantly different such that a different approach is required.

However, text or document classification, the general technique employed in this paper, is a very common task (Manning et al., 2008). In text classification, given a document d, the task is to assign it a class, or label, c, from a fixed, human-defined set of possible classes $C = \{c_1, c_2, \ldots, c_n\}$. In order to achieve this, a training set of *labelled documents* $\langle d, c\rangle$ is given to a *learning algorithm* to learn a classifier that maps documents to classes.

Documents are typically represented as a vector in a high-dimensional space, such as term-document matrices, or results of dimensionality reduction techniques such as Latent Semantic Analysis (Landauer et al., 1998), or more recently, using vector representations of words produced by neural networks (Pennington et al., 2014).

Text classification has many applications, among others spam filtering (Androutsopoulos et al., 2000) and sentiment analysis (Pang and Lee, 2008).

One particular application that could be deemed of relevance with respect to our work is that of readability assessment (Pitler and Nenkova, 2008; Feng et al., 2010), i.e. determining the ease with which a written text can be understood by a reader, since age is certainly a dimension along which readability varies. However, our literature review of this area suggested that the aspects being considered in readability assessment are sufficiently different from the dimensions that seem to be most relevant for media age appropriatness ratings. Following Manurung et al. (2008), we hypothesize that utilizing resources such as the MRC Psycholinguistic Database (Coltheart, 1981) could be valuable in determining age appropriateness, in particular various features such as familiarity, imageability, age-of-acquisition, and concreteness.

3 Corpus Construction

There are three steps in obtaining the data required for our corpus: obtaining album details and age-appropriateness ratings, searching for the tracklisting of each album, and obtaining the lyrics for each song. Each step is carried out by querying a different website. To achieve this, a Java application that utilizes the jsoup library[4] was developed.

3.1 Obtaining album details and age-appropriateness ratings

The Common Sense Media website provides reviews for various music albums. The reviews consist of a textual review, the age-appropriate rating for the album, which consists of an integer in the interval [2,17] or the label 'Not For Kids', and metadata about the album such as title, artist, and genre. Aside from that, there are also other annotations such as a quality rating (1-5 stars), and specific aspectual ratings such as positive messages, role models, violence, sex, language, consumerism, drinking, drugs & smoking. The website also allows visitors to contribute user ratings and reviews. In our experiments we only utilize

[4] http://www.jsoup.org

the album metadata and integer indicating the age-appropriate rating.

3.2 Tracklist searching

A tracklist is a list of all the songs, or tracks, contained within an album. From the information previously obtained from Common Sense Media, the next step is to obtain the tracklist of each album. For this we query the MusicBrainz website[5], an open music encyclopedia that makes music metadata available to the public. To obtain the tracklists we employed the advanced query search mode that allows the use of boolean operators. We tried several combinations of queries involving album title, singer, and label information, and it turned out that queries consisting of album title and singer produced the highest recall. When MusicBrainz returns multiple results for a given query, we simply select the first result. For special cases where the tracks on an album are performed by various artists, e.g. a compilation album, or a soundtrack album, it is during this stage that we also extract information regarding the track-specific artist name. Finally, we assume that if the album title contains the string 'CD Single' then it only contains one track and we skip forward to the next step.

3.3 Lyrics searching

For this step, we consulted two websites as the source reference for song lyrics, songlyrics.com and lyricsmode.com. The former is first consulted, and only if it fails to yield any results is the latter consulted. If a track is not found on both websites, we discard it from our data set. Similar to the previous step, we perform a query to obtain results, however during this step the query consists of the song title and singer. Once again, given multiple results we simply choose the first result. In total, we were able to retrieve lyrics from 12,242 songs across 1,798 albums. Table 1 provides an overview of the number of tracks and albums obtained per age rating.

4 Experimentation

Since the constructed data set is imbalanced, we use the SMOTE oversampling technique to overcome this problem (Chawla et al., 2002). This results in a balanced dataset with the same number of samples in each class.

[5]http://www.musicbrainz.org

Group	Age	#Tracks	#Albums
Toddler	2	696	119
	3	130	23
Pre-schooler	4	251	46
	5	204	31
Middle childhood 1	6	281	41
	7	358	71
	8	654	118
Middle childhood 2	9	237	50
	10	1,590	253
	11	580	105
Young teen	12	1,849	253
	13	1,767	242
	14	1,453	177
Teenager	15	653	116
	16	521	64
	17	180	16
Adult	>17	838	73
	Total	12,242	1,798

Table 1: Statistics of the dataset

Once the dataset is complete, classifiers were trained and used to carry out experiment scenarios that vary along several factors. For the class labels, two scenarios are considered: one where each age rating from 2 to 17 and 'Not For Kids' is a separate class, and another where the data is clustered together based on some conventional developmental age groupings[6], i.e. toddlers (ages 2 & 3), preschoolers (ages 4 & 5), middle-childhood 1 (ages 6 to 8), middle-childhood 2 (ages 9 to 11), young-teens (ages 12 to 14), and teenagers (ages 15 to 17), with an additional category for ages beyond 17 using the 'Not For Kids' labelled data.

For the instance data, two scenarios are also considered: one where classification is done on a per-track basis, and one on a per-album basis (i.e. where lyrics from all its constituent tracks are concatenated).

As for the feature representation, three primary variations are considered:

Vector Space Model. This is a baseline method where each word appearing in the dataset becomes a feature, and a vector representing an instance consists of the $tf.idf$ values of all words. Additionally, stemming is first performed on the words, and information gain-based attribute selection is applied.

MRC Psycholinguistic data. For this feature

[6]http://www.cdc.gov/ncbddd/childdevelopment/positiveparenting/

representation, given each distinct word appearing in the lyrics of a track (or album), a lookup is performed on the MRC psycholinguistic database, and if appropriate values exist, they are added to the tally for the familiarity, imageability, age-of-acquisition, and concreteness scores. Thus, an instance is represented by a vector with four real values. The vectors are normalized with respect to the number of words contributing to the values.

GloVe vectors. GloVe[7] is a tool that produces vector representations of words trained on very large corpora (Pennington et al., 2014). It is similar to dimensionality reduction approaches such as latent semantic analysis. For this experiment, the 50-dimensional pre-trained vectors trained on Wikipedia and Gigaword corpora were used.

When combining feature representations, we simply concatenate their vectors.

Finally, for the classification itself, the Weka toolkit is used. Given the ordinal nature of the class labels, classification is carried out via regression (Frank et al., 1998), using the M5P-based classifier (Wang and Witten, 1997). The experiments were run using 4-fold cross validation.

For the initial experiment, only the baseline VSM feature representation was used, and the treatment of class labels and instance granularity was varied. The results can be seen in Table 2, which shows the average accuracy, i.e. the percentage of test instances that were correctly labelled, across 4 folds.

	Age group	Year
Per-track	69.77%	58.58%
Per-album	70.60%	57.15%

Table 2: Initial experiment varying class and instance granularity

For the follow-up experiment, we focus on the task of classifying at the per-album level of granularity, as ultimately this is the level at which the original annotations are obtained. For the class labels, both age groups and separate ages are used. The feature representation was varied ranging from VSM, VSM + MRC, VSM + GloVe, and VSM + GloVe + MRC. The results can be seen in Table 3.

[7]http://nlp.stanford.edu/projects/glove/

Features	Age group	Year
VSM	70.60%	57.15%
VSM + MRC	**71.02%**	56.80%
VSM + GloVe	70.58%	57.68%
VSM + GloVe + MRC	70.47%	**57.85%**

Table 3: Results varying feature representations

5 Discussion & Summary

From the initial experiment, it appears that distinguishing tracks at the level of granularity of specific year/age (e.g. "is this song more appropriate for a 4 or 5 year old?") is very difficult, as indicated by an accuracy of only 57% to 58%. Bear in mind, however, that this is a seventeen-way classification task. Shifting the level of granularity to that of age groups transforms the task into a more feasible one, with an accuracy around the 70% mark. It is surprising to note that the per-track performance is better than the per-album performance when tracks are distinguished by specific age/year rather than age groups. We had initially hypothesized that classifying albums would be a more consistent task given the increased context and evidence available.

As for the various feature representations, we note that the addition of the MRC psycholinguistic features of familiarity, imageability, concreteness, and age-of-acquisition does provide a small accuracy increase in certain cases, as evidenced by the highest accuracy of 71.02% when classifying albums by age group using the VSM + MRC features. The use of the GloVe vectors gives a slight contribution in the case of classifying albums by specific age/year, where the highest accuracy of 57.85% is obtained when combining VSM with both the MRC and GloVe features.

There are many other features and contexts that can also be utilized. For instance, given the metadata of artist, album, and genre, additional information may be extracted from the web, e.g. the artist's biography, general-purpose album reviews, genre tendencies, etc., all of which may contribute to discerning age-appropriateness. Another set of features that can be utilized are readability metrics, as they are often correlated with the age of the reader.

To summarize, this paper has introduced a novel task with clear practical applications in the form of automatically identifying age-appropriate ratings of songs and albums based on lyrics. The work

reported is still in its very early stages, nevertheless we believe the findings are of interest to NLP researchers.

Another question that needs to be addressed is what sort of competence and agreement humans achieve on this task. To that end, we plan to conduct a manual annotation experiment involving several human subjects, themselves varied across different age groups, and to measure inter-annotator reliability (Passonneau et al., 2006).

References

Ion Androutsopoulos, John Koutsias, Konstantinos Chandrinos, Georgios Paliouras, and Constantine D. Spyropoulos. 2000. An evaluation of naïve Bayesian anti-spam filtering. In *Proceedings of the workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning*, pages 9–17, Barcelona, Spain.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, June.

Max Coltheart. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33(4):497–505.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 276–284, Stroudsburg, PA, USA. Association for Computational Linguistics.

E. Frank, Y. Wang, S. Inglis, G. Holmes, and I.H. Witten. 1998. Using model trees for classification. *Machine Learning*, 32(1):63–76.

Thomas Landauer, Peter Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Christopher Manning, Prabhakar Raghavan, and Hinrich Schutze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Ruli Manurung, Graeme Ritchie, Helen Pain, Annalu Waller, Dave O'Mara, and Rolf Black. 2008. The construction of a pun generator for language skills development. *Applied Artificial Intelligence*, 22(9):841–869.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Rebecca Passonneau, Nizar Habash, and Owen Rambow. 2006. Inter-annotator agreement on a multilingual semantic annotation task. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, May.

Jeffrey Pennington, Richard Socher, and Christopher 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. Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 186–195, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. Polpinij, A. Chotthanom, C. Sibunruang, R. Chamchong, and S. Puangpronpitag. 2006. Content-based text classifiers for pornographic web filtering. In *Systems, Man and Cybernetics, 2006. SMC '06. IEEE International Conference on*, volume 2, pages 1481–1485, Oct.

Sara Owsley Sood, Judd Antin, and Elizabeth F Churchill. 2012. Using crowdsourcing to improve profanity detection. In *AAAI Spring Symposium: Wisdom of the Crowd*.

Gui-yang Su, Jian-hua Li, Ying-hua Ma, and Shenghong Li. 2004. Improving the precision of the keyword-matching pornographic text filtering method using a hybrid model. *Journal of Zhejiang University Science*, 5(9):1106–1113.

Y. Wang and I. H. Witten. 1997. Induction of model trees for predicting continuous classes. In *Poster papers of the 9th European Conference on Machine Learning*. Springer.

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*, CIKM '12, pages 1980–1984, New York, NY, USA. ACM.