Predict Box-office Success Based on Textual, Visual, and Movie Metadata with Various Machine Learning Models

Wen Xie and Yigeng Zhang
University of Houston
{wxie5, yzhang168}@uh.edu

Abstract

Movies are popular entertainment loved by people all over the world. Considering the investment and the market, predicting the likeability of a movie is always interesting to the movie makers. In this work, we propose two machine learning-based methods to predict the likeability of a movie based mainly on the subtitles and metadata. By investigating both traditional machine learning techniques and deep learning models, we achieved a comparable good result compared to the previous baselines.

1 Introduction

The audience likeability is fundamental to the success of a movie. Getting the opinions of the movie product from the target customers is always desired by the providers because this information is helpful for them to analyze the market and prepare for future production. However, collecting the opinions from a large number of users in a traditional way is almost impossible considering the workload.

Automatically predicting the movie likability using a large amount of data now draws the interests of researchers and businesses. Recently, there are much user-generated content regarding the rating in terms of a movie. There is also much metadata related to a movie product such as actors, directors, and movie genres, which are easy to obtain from the Internet. Meanwhile, the subtitle and the ploy of the movies are also informative data reflecting the movie content. By making use of various data, people start their attempt to investigate the success of a movie using data-driven empirical methods.

In this study, we propose to use machine learning-based NLP approaches to predict the audience likeability of movies based on their subtitles, metadata, and posters. Our contribution in this work can be summarized as follow:

- We propose various machine learning techniques to solve the movie likability prediction problem and achieved comparable results to the previous work.
- We introduced and investigated different new features such as visual attention and BERT sentence embedding to extend the previous baseline.

2 Related Work

2.1 Movie Content Analysis

In recent years, research on movie content analysis has been popular. Movie content analysis is usually based on visual, audio, and textual data. Many previous efforts focused on violent scene detection. (Datta et al., 2002) used visual information in violence scene detection, while Rasheed, et al. (Rasheed and Shah, 2002), and Giannakopoulos, et al.(Giannakopoulos et al., 2006)(Giannakopoulos et al., 2007) used both audio and video information in this task. These works are based on pre-designed feature selection and traditional machine learning techniques such as SVM.

Comparing to the work on visual and audio information mentioned above, Kar, et al. (Kar et al., 2018) make use of plot synopses to create descriptive tags for movies. Shafaei, et al. (Shafaei et al., 2019b) were the first to propose to predict the MPAA ratings for movies. They predict the four-level ratings by taking the script data using a deep neural network method. More specifically, Martinez, et al. (Martinez et al., 2019) analyzed the language usage in scripts to characterize the levels of violence in movies. From the likability prospective, (Shafaei et al., 2019a) proposed a large dataset with movie scripts, posters, and metadata. They built up a strong baseline using comprehensive features to predict the likability of a movie.

2.2 Document Representation and Classification

Since the major part of data in this task is the subtitles. Predicting the binary or multi-level likability is a text classification problem. Traditional text processing methods such as Bag-of-Words (Joachims, 1998)(McCallum et al.), TF-IDF (Luhn, 1957)(Jones, 1972) present with simple yet effective statistical text representation for traditional machine learning techniques. In recent years, deep learning methods take a dominant place in solving text classification problems. The advantage of neural network-based models is that they can extract features and learning representations of the data automatically. Some prior works such as (Kim, 2014) and (Joulin et al., 2017) shows the power of deep learning in text classification. Recently, Transformer (Vaswani et al., 2017) based models shows promising results in many NLP tasks (Devlin et al., 2018a). Meanwhile, good-quality representations of long documents play an important part in the final classification process while the previous works do well mostly limited on sentences or short passages(Kim, 2014)(Akbik et al., 2019). This is one of the challenging and essential parts of the task.

In this work, we also focus on the script data with metadata and visual information to predict the likability of the movies. We treat the task as a text classification/regression problem and apply various traditional machine learning and deep learning methods.

3 Dataset

3.1 Dataset Basics

Mainly, we use the dataset offered in the movie project provided in the course. Briefly speaking, the dataset includes metadata and subtitles of movies collected from the Internet. The metadata includes around 16,000 movies in total, around 5,000 of which have the box office revenue information. The subtitles include more than 22,000 movie scripts. The metadata in the dataset includes the movie's basic information such as the directors, cast, movie genre, and run times. Table 1 summarizes the counts of each movie genre type. Figure 1 shows the distributions of movie release year and runtime length.

Genre	Count	Genre	Counts
Sci-Fi	888	Fantasy	110
Action	2533	Thriller	2227
Horror	946	Short	129
Animation	601	Mystery	960
Crime	1274	Film-Noir	191
History	513	Music	399
Romance	2415	Drama	3474
Adventure	1029	Documentary	805
News	19	Family	112
Western	218	Sport	298
Comedy	3535	Biography	592
War	398		

Table 1: Movie genre statistics.

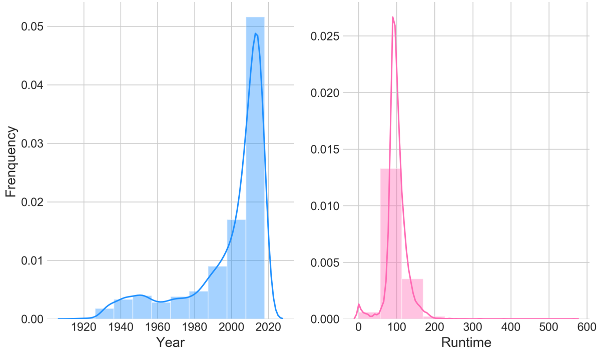

Figure 1: Release year and runtimes distribution

To make classification and regression, we consider three kinds of features including movie product information, visual attention distribution, and scripts textual information. In order to keep model input features consistent, we delete the movies whose poster links are absent. We also delete movies whose scripts size is less than 5B. Finally, we have 15195 movies.

3.2 Success Criteria

Similarly to previous study (Shafaei et al., 2019a), we also take IMDB rating as movie success criteria. Figure 2 show the IMDB rating distribution. Clearly, most movies receive ratings from 5 to 8. In binary classification case, IMDB rating is equal or larger than 6.5 is regarded as success. Movie whose IMDB rating is less than 6.5 is a failure. In multi-class classification case, we add one more label to fine the classes. Movie whose IMDB rating is equal or larger than 7 is successful product while movie with less than 6.5 IMDB rating is taken as a failure product. Movies with the middle range from 6.5 to 7 rating are average products. Table 2 shows the basic summary information. In regression case, the raw IMDB rating value is the ground-truth.

Binary	Count	Multiclass	Count
Successful	7,084	Successful	4,398
Failure	8,111	Average	5,271
		Failure	5,526
Total	15,195	Total	15,195

Table 2: Classification labels.

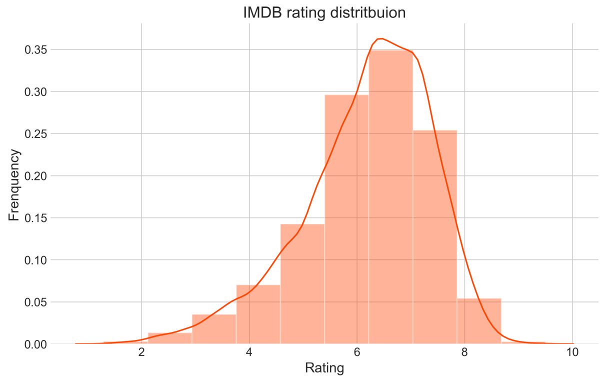

Figure 2: IMDB rating distribution

4 Methodology

This research problem is primarily a classification/regression task on multimodal data. Therefore, typical procedures like feature extraction, model development, and evaluation will be conducted.

4.1 Feature Extraction

In this work, we have mainly three types of features: textual features, visual features, and product features.

4.1.1 Traditional Features

Genre, Actors, Director Traditionally, researchers use movies' metadata to predict box-office success. In this study, we select genres, actors, and directors (Gmerek et al., 2015). Each variable is encoded into a one-hot vector.

Runtime, Release Year Besides genre, actor, and director, runtime and the release year are also considered. We assume that movies with different lengths are likely to receive different preferences. Movies released in different years also may have a big difference in their success. For example, the national policy, macro-economy may make a difference to the motion picture industry. Impacted by the COVID-19, all movie theaters are shut down during the first half of 2020. Movie runtime and release year variables are coded into OneHot vector.

4.1.2 Proposed Features

NRC Word-Emotion Association lexicon Textual data are the movie scripts, which are in the form of many dialogue subtitles. It contains much information but not necessarily to be the most contributive feature. Besides the text embeddings, we extract the word-level sentimental and emotional features by NRC Word-Emotion Association lexicon (Mohammad and Turney, 2013) as the prior work did (Shafaei et al., 2019a). We use the NRC Word-Emotion Association lexicon version 0.92, which includes 14,182 unigrams (words), and around 25,000 word senses. The sentiments include negative and positive manually annotated labels by the crowdsourcing website Amazon's Mechanical Turk. The word emotions include eight categories: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. Considering the development of a movie storyline, each script is equally divided into five sections. We first use pre-trained BERT tokenizer (Devlin et al., 2018b) to process each sentence of each script. Then, we look up the association of each word in each section. At last, the total associations of all sentences in each section are used to reflect the word-level sentiments and emotional features of the section. The emotional and sentimental features of each script are measured by the five sections together.

Writing Style We extract many morphology and word shape features as used in (Maharjan et al., 2017)'s paper to indicate the potential dissimilar usage in successful and unsuccessful writings. In summary, we consider the following features: the number of words, characters, uppercase words, title words, exclamations, question marks, the average word length, sentence length, words per sentence, and lexical diversity of each script.

Sentence Embedding In recent years, many state-of-art pre-trained embedding networks have been proposed such as BERT (Devlin et al., 2018a) and RoBERTa (Liu et al., 2019). However, These models require a lot of computation and memory when facing high dimensional input data. In this study, we extract contextual sentence embedding features using a pre-trained BERT network built on siamese and triplet structures which are proposed by (Reimers and Gurevych, 2019). When finding the semantically meaningful sentence embeddings, the siamese and triplet networks take much less time compared to the original BERT network based on deep bi-direction transformer structure. The output dimension of the pre-trained BERT model is 768. Considering the limited computation ability, we furthermore lower the dimension to 200 by Principal Component Analysis(PCA).

Visual Features We consider movies' posters as visual features. Compared with textual data, images deliver information directly and are likely to attract more attention. Based on the poster link of each movie in the metadata, we download the movie posters and extract visual features from the posters. Instead of extracting a latent feature from an image processing network trained on general image datasets such as ImageNet, we choose Convolutional Neural Networks (CNN) trained on eye-fixation labeled images (Fan et al., 2018). When browsing movie posters, people usually show potential consistent characteristics. For example, the image attribute in the poster is likely to attract more attention compared to text information. We propose that CNN trained on eye-fixation labeled images can capture the attention distribution feature more completely and usefully.

4.2 Traditional Methods

In this project, we consider both traditional tree-based algorithms and neural networks for the classification and regression tasks. In terms of traditional algorithms, we mainly consider the Support Vector Machine (SVM) and LightGBM.

SVM is well known and performs excellently in many regression and classification tasks especially when facing homogeneous training and testing datasets. Kernel method in SVM is also a very useful technique that is able to map data inseparable in low dimension space to separable data in high dimension space. However, we also choose lightGBM (Ke et al., 2017) which is a gradient boosting framework that uses tree-based learning algorithms. LIghtGBM is designed to be distributed and efficient with the following advantages: a faster training speed and higher efficiency; lower memory usage; support of parallel and GPU learning; capable of handling large-scale data. Because of the above-mentioned advantages, it's suitable for our tasks given our big movie dataset.

The traditional approach framework is shown in Figure 3. After data preprocessing and feature extraction, we feed features into SVM and lightGBM model to finish classification and regression tasks.

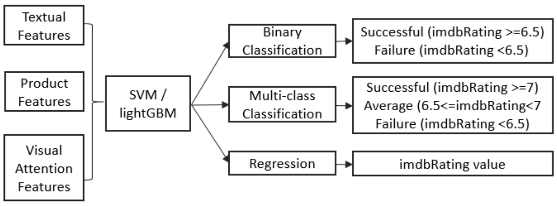
Textual Features
Product Features
Visual Attention Features
SVM / lightGBM
Binary Classification
Multi-class Classification
Regression
Successful (imdbRating >=6.5)
Failure (imdbRating <6.5)
Successful (imdbRating >=7)
Average (6.5<=imdbRating<7
Failure (imdbRating <6.5)
imdbRating value

Figure 3: SVM and lightGBM tasks

4.3 Deep Learning Methods

Neural-network-based models (such as RNNs and Transformers) have shown power in many NLP tasks in recent years. In this work, we implement a deep learning pipeline focusing on investigating the influence of the subtitle text. Hierarchical Attention Networks (HAN) (Yang et al., 2016) is proven to be an effective text classification method. This model intuitively applies the attention mechanism on the word level and sentence level hierarchically, which therefore makes it possible to capture the informative element throughout the document.

The word-level encoder consists of a bi-directional GRU with attention, which is described in the following equations, where w represents word embedding and T is the number of words in the sentence, with $t \in [1, T]$ forward, or $t \in [T, 1]$ backward.

$$\begin{aligned} \vec{h}_{it} &= \overrightarrow{\mathrm{GRU}}\left(w_{it}\right) \\ \overleftarrow{h}_{it} &= \overleftarrow{\mathrm{GRU}}\left(w_{it}\right) \end{aligned} \tag{1}$$

$$\begin{aligned} u_{it} &= \tanh\left(W_w h_{it} + b_w\right) \\ \alpha_{it} &= \frac{\exp\left(u_{it}^{\top} u_w\right)}{\sum_t \exp\left(u_{it}^{\top} u_w\right)} \\ s_i &= \sum_t \alpha_{it} h_{it} \end{aligned} \tag{2}$$

The sentence-level encoder and attention work in a similar way, where s is sentence representation learned from the previous step and N is the number of sentence in the document, with $i \in [1, N]$ forward, or $i \in [N, 1]$ backward.

$$\begin{aligned} \vec{h}_{i} &= \overrightarrow{\mathrm{GRU}}\left(s_{i}\right) \\ \overleftarrow{h}_{i} &= \overleftarrow{\mathrm{GRU}}\left(s_{i}\right) \end{aligned} \tag{3}$$

$$\begin{aligned} u_{i} &= \tanh\left(W_s h_{i} + b_s\right) \\ \alpha_{i} &= \frac{\exp\left(u_{i}^{\top} u_s\right)}{\sum_i \exp\left(u_{i}^{\top} u_s\right)} \\ v &= \sum_i \alpha_{i} h_{i} \end{aligned} \tag{4}$$

The model structure is illustrated as follows:

4.4 Evaluation Criteria

We first define the concept of 'likability' or 'success' in this work as a numerical value. In the previous work (Shafaei et al., 2019a), the authors scale down the concept into binary (Successful/Unsuccessful) or ternary (Successful/Average/Unsuccessful) level according to the

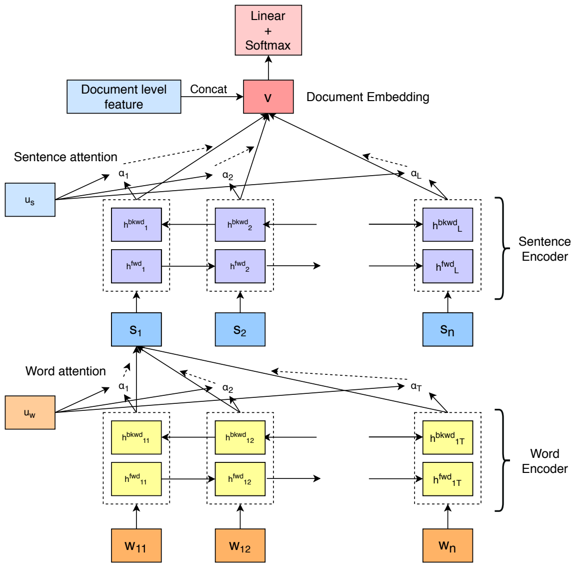

Figure 4: Model structure of Hierarchical Attention Networks.

	Predicted class	
Actual class	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Table 3: True/false positive/negative explained.

IMDB rating score with a threshold. To make our work comparable to the previous work, we make our ‘likability’ criteria in line with the previous one and keep the same classification labels.

From the classification perspective, the F1 score can be the major metric for evaluation. From the regression perspective, measuring how close the prediction is to the true score should be the key criterion.

One of the most important evaluation metrics in this type of task is F1 score. To introduce precision, recall, and F1 score, we need to first figure out true/false positive/negative indicators, which is illustrated in Table 3. Precision is the ratio of correctly predicted positive labels to the total predicted positive labels. Recall is the ratio of correctly predicted positive labels to the total existing positive labels. F1 score is a weighted average of Precision and Recall, which makes this metric works better in evaluating uneven label distributions.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

5 Experiment Results

We first pre-process the multimodal data and then extract all the proposed features. Then we implement the traditional machine learning models (SVM and lightGBM) and Deep learning model (Hierarchical Attention Network) for the classification and regression tasks. After that, we will conduct experiments iteratively and evaluate the result. Finally, we conduct some correlation analysis between the movie’s metadata and box-office.

5.1 Results of Traditional Methods

In total, our dataset includes 15,195 movie samples. The movie samples are randomly divided the movie samples into training and testing samples in the proportion of 80:20. Then, SVM and lightGBM models are utilized to do binary classification, multi-class classification, and regression. The performances of the two models in each task are documented in Table 4. Because of our hardware limitations, we were not able to perform experiments on features like unigram or bigram.

Feature	**BC(f1)**	**MC(f1)**	**RE(mse)**
NRC emotion(1)	0.68	0.53	1.38
Writing style(2)	0.69	0.52	1.45
Sents embedding(3)	0.61	0.44	1.35
Visual attention(4)	0.69	0.54	1.42
Release year(5)	0.64	0.48	1.30
Runtime(6)	0.63	0.47	1.31
Genre(7)	0.66	0.49	1.10
Director(8)			
Actors(9)			
1,2,3,4	0.69	0.52	1.42

Table 4: Performances of SVM.

Table 4 and Table 5 record the F1 score and MSE of SVM and lightGBM models in each task. Due to time limits, computer memory, and computation ability limitations, we’re not allowed to conduct more experiments and report more reliable results. The performances are basic results of SVM and lightGBM models without finetuning.

Feature	BC(f1)	MC(f1)	RE(mse)
NRC emotion(1)	0.60	0.49	1.38
Writing style(2)	0.69	0.52	1.42
Sents embedding(3)	0.60	0.42	1.35
Visual attention(4)	0.59	0.42	1.40
Release year(5)	0.61	0.47	1.29
Runtime(6)	0.63	0.46	1.29
Genre(7)	0.63	0.48	1.17
Director(8)	0.69	0.54	1.42
Actors(9)	0.69	0.53	1.42
1,2,3,4	0.59	0.42	1.36

Table 5: Performances of lightGBM.

5.2 Results of Deep Learning Methods

We first use only the text of subtitle, the HAN model achieves a weighted F1 score 0.7006 on binary classification, which has already outperformed the performance of all of the traditional methods with feature selection. Then we added a single feature, movie genre, at the document representation level. The experimental result of using HAN is listed below.

Feature	BC(f1)	MC(f1)
Text + Genre	0.7241	0.5475

Table 6: Performances of Hierarchical Attention Network.

The performance improved 0.024 points in the binary classification performance and achieved 0.5475 in the multi-class classification. This result is the best among all of the methods we have experimented with. Due to the limitation of time and hardware, we are not able to conduct more experiments using deep learning methods.

6 Analysis and Discussion

On binary classification, the experimental results on traditional methods show the models using visual attention have the best performance. We hypothesize that the visual information on the posters could reflect the attributes of actors and special effects, which make it informative in judging the likability by users. NRC emotion also contributes to the classification. The proposed method did not have visible improvement using simply the BERT sentence embedding. The reason could be that sentence embedding is less informative if the model focuses more on the word level features as it encodes a more abstract level of the feature. This is evidenced by the previous work as the n-gram models presented a very strong baseline.

The performance of the multi-class task is much lower than the binary. The best performance also comes from the visual information on the posters, which makes us curious to make it as future exploration. The director and actor information also gives a good performance, from which we hypothesize that the reputation of the actors and directors is in line with the likability of a movie to some extent.

For the model perspective, we can easily find that the performances vary on different traditional methods, even if they perform classification and regression on the same set of features. From the results it is difficult to distinguish which model is better, however, the lightGBM model runs much faster than SVM, which is a significant advantage. The deep learning model easily beats the traditional model using only the textual data. The HAN model makes the most of the text throughout by using attention from word level to sentence level and finally achieved the best result among all the methods with feature combinations. It is promising to have more improvement if we add more features.

7 Conclusion

In this work, we did a throughout the study on movie likability prediction. By focusing on textual subtitle data along with metadata and poster data, we achieved comparable performance to the previous work using different machine learning techniques. The experimental result showed the effectiveness of the methods we developed and gave prominence to the power of deep learning models.

Because of the limitation of our hardware and time, we are not able to investigate more on the deep learning methods. (Each experiment on HAN takes several hours). We are optimistic that with more time for hyper-parameter selection and more metadata incorporated into the deep learning approach, there is still space for performance improvement.

For future work, we plan to introduce more potential helpful features such as visual information and metadata of the actors and directors. We can also improve the current method with transfer learning methods such as pre-training our deep model on a larger dataset such as general text.

Acknowledgments

We would like to thank Prof. Solorio for this great course and we both learned a lot during this semester. We also want to thank our TA Niloofar and Mahsa for the instructions and help in the class and homework.

A Appendices

A.1 Contribution of the authors

The authors have equal contributions to this work. Wen focused on traditional methods implementation and evaluation. Yigeng focused on deep learning methods implementation and evaluation. They discuss and help each other in their coding.

They finished this report together and put efforts equally. Wen is mainly responsible for section 3, 4.1, 4.2, 5.1 Yigeng is mainly responsible for section 1, 2, 4.3, 4.4, 5.2. They helped mutually with each other in their responsible sections in writing and they discussed and finished the analysis and conclusion sections together.

References

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Pooled contextualized embeddings for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 724–728.

Ankur Datta, Mubarak Shah, and N Da Vitoria Lobo. 2002. Person-on-person violence detection in video data. In *Object recognition supported by user interaction for service robots*, volume 1, pages 433–438. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L Koenig, Juan Xu, Mohan S Kankanhalli, and Qi Zhao. 2018. Emotional attention: A study of image sentiment and visual attention. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 7521–7531.

Theodoros Giannakopoulos, Dimitrios Kosmopoulos, Andreas Aristidou, and Sergios Theodoridis. 2006. Violence content classification using audio features. In *Hellenic Conference on Artificial Intelligence*, pages 502–507. Springer.

Theodoros Giannakopoulos, Aggelos Pikrakis, and Sergios Theodoridis. 2007. A multi-class audio classification method with respect to violent content in movies using bayesian networks. In *2007 IEEE 9th Workshop on Multimedia Signal Processing*, pages 90–93. IEEE.

Natalia Gmerek et al. 2015. The determinants of polish movies' box office performance in poland. *Journal of Marketing and Consumer Behaviour in Emerging Markets*, 1(1):15–35.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Armand Joulin, Edouard Grave, and Piotr Bojanowski Tomas Mikolov. 2017. Bag of tricks for efficient text classification. *EACL 2017*, page 427.

Sudipta Kar, Suraj Maharjan, and Thamar Solorio. 2018. Folksonomication: Predicting tags for movies from plot synopses using emotion flow encoded neural network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2879–2891.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pages 3146–3154.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Hans Peter Luhn. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317.

Suraj Maharjan, John Arevalo, Manuel Montes, Fabio A González, and Thamar Solorio. 2017. A multi-task approach to predict likability of books. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1217–1227.

Victor R Martinez, Krishna Somandepalli, Karan Singla, Anil Ramakrishna, Yalda T Uhls, and Shrikanth Narayanan. 2019. Violence rating prediction from movie scripts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 671–678.

Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. Citeseer.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

Zeeshan Rasheed and Mubarak Shah. 2002. Movie genre classification by exploiting audio-visual features of previews. In *Object recognition supported by user interaction for service robots*, volume 2, pages 1086–1089. IEEE.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Mahsa Shafaei, Adrián Pastor López-Monroy, and Thamar Solorio. 2019a. Exploiting textual, visual, and product features for predicting the likeability of movies. In *The Thirty-Second International Flairs Conference*.

Mahsa Shafaei, Niloofar Safi Samghabadi, Sudipta Kar, and Thamar Solorio. 2019b. Rating for parents: Predicting children suitability rating for movies based on language of the movies. *arXiv preprint arXiv:1908.07819*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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.