Prediction of book rating and genres by machine learning algorithms

Sena Baskan

Sabanci University, Data Analytics Master Program
Istanbul, Turkey

Keywords:
Natural Language Processing
Text Mining
Topic Modelling
Image Processing

A B S T R A C T

Amount of unstructured data created by computers is increasing every day with an incredible growth rate. It is a challenging problem to analyze unstructured data - video, image, text, audio, etc. Latent information that can be extracted from these type of data gets more and more valuable every day as they provide both academical and commercial organizations with crucial insight and vision for their study and activities. Machine learning algorithms are utilized to transform such type of data to an understandable and interpretable state. This paper presents applications of machine learning methods on text ('book description, title, genre') and image ('book cover') data of books. There are two parts of this study. The aim of the first part is to find the best rating prediction model by using text mining and afterwards regression algorithms. Second part aims to predict genres of the books by applying color identification and then classification on the extracted features.

1. Introduction

For many of the organizations, information that can be derived from data is crucial to make right decisions and continue their existence in traditional economic system. Most of them need to analyze text and image data. Text mining handles text data and extracts relevant information. User profiling and recommendation for social media and e-commerce, customer relationship management, spam filtering can be given as examples that text mining leverages. Image processing has wide application areas like astronomy, medical field, aviation, robotics, etc.

Publishing companies, that publish new books every year, surely have to deal with massive amount of texts. Estimation of a candidate book's commercial success is important for them to decide if it is going to be published or not. A great number of long documents should be read and this requires manpower. For this reason, an automatic evaluation system which utilizes machine learning methods can be significantly valuable for publishers. Besides, cover image which forms first impression on the potential readers, can be analyzed by image processing to give idea to the designers.

In this study, "Best Books Ever" data[1] gathered from 'GoodReads' - an international community platform on books – is analyzed. Dataset comprises book description as a summary of the book, author name, book title, genre as text data; cover image as image data; other numerical attributes and average book rating given by Goodreads users. The following questions are the focus of the study:

1. "Can a rating prediction model be found out by using text and numerical data belonging to books?"
2. "Are book title and cover image color distribution factors that can help to identify the genre of the book?"

This study is an experimental one. As far as it has been researched this study seems to be the first attempt to experiment:

- usage of topic modelling to predict book rating.
- usage of simple Red Green Blue (RGB) color distribution together with book title to predict genre of the book.

Most of the studies on rating predictions related with text are based on sentiment analysis of user reviews. This study takes into account not subjective expressions, objective book information in text form. This study is based on text decomposition, not classification.

An overview of related works is given in Section 2. Methodology details and results are given in Section 3. Conclusions on results are presented finally.

[1] The data can be downloaded from https://www.kaggle.com/meetnaren/goodreads-best-books

2. Related Work

2.1. Text Mining

Text mining tries to convert unstructured text data to information through the identification of patterns in a collection of documents. Clustering is an approach in text mining that provides a transformation of the data that leads up to extract/create features as inputs to the machine learning algorithms which need well defined fixed length inputs. There are several studies on text mining. Some of them are as following.

Mooney, Bennett and Roy (1998) developed a book recommendation system that utilizes semi-structured information about items gathered from the web using simple information extraction techniques.

Hotho, Nürnberger and Paaß (2005) gave a brief introduction to the broad field of text mining and presented overview of available text mining methods, their properties and their application to specific problem.

Afonso and Duque (2014) tried to verify whether an automated clustering process could create the correct clusters for scientific and newspaper corpuses in different languages.

Lydia, Govindaswamy, Lakshmanaprabu and Ramya (2018) worked on detailed working process involved in the K-means algorithm and in document preprocessing.

Hadifar, Sterckx, Demeester and Develder (2019) proposed a method that learns discriminative features from both an autoencoder and a sentence embedding, then uses assignments from a clustering algorithm as supervision to update weights of the encoder network.

2.2. Topic Modelling

Topic models are probabilistic models that uncover the hidden structure of text corpus and decompose documents into topics. Words are assumed to belong to those topics. One of the most popular methods to apply topic modelling is Latent Dirichlet allocation (LDA) developed by Prof. David M. Blei in 2003. Some of the studies on topic modelling are as following.

Blei, Ng, and Jordan (2003) described Latent Dirichlet allocation (LDA) model, a three-level hierarchical Bayesian model, in which each item of a collection of data is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities.

Alikaniotis, Yannakoudakis and Rei. (2003) introduced a deep neural network model capable of representing both local contextual and usage information as encapsulated by essay scoring.

Wallach (2006) showed that a model that integrates bigram-based and topic-based approaches to document modeling has several benefits.

Wang, McCallum, Wei (2007) proposed a topical n-gram model that determines to form an n-gram or not based on context and assign mixture of topics to both individual words and n-gram phrases.

Tong and Zhang (2016) implemented experiments on Wikipedia articles and Twitter tweets by topic modelling and designed a computational tool.

Arseniy and Natalia (2016) proposed a modification of the Anchor Words topic modeling algorithm that takes into account collocations. They showed that this approach leads to the increase of the interpretability without deteriorating perplexity.

Some of the studies that used text mining and regression algorithms like this study are as following.

Maharjan, Gonzales, Montes-y-G´omez, Solorio (2017) proposed new features for predicting the success of books. They predicted book ratings by using only the contents of the books.

An anonymous post on Kaggle published a study on book rating prediction by linear regression method, taking features as label encoded author and title and book pages review count.

2.3. Image Processing

Image processing which has extensive applications in many areas, like medicine, industrial robotics, and remote sensing by satellites, is a set of computational techniques on an image, in order to get an enhanced form it or to analyze and extract some useful information from it. Its main components are importing, that is capturing the image, analysis and manipulation by specialized software applications and acquiring the output. Some of the studies on image processing are as following.

Zujovic, Gandy, Friedman, Pardo, Pappas (2009) studied on painting genre detection. They tackled the problem by extracting features from gray scale images. They then tried multiple different classifiers.

Kulkarni, Kurundkar, Khare, Savant, Chintal (2015) reviewed development and implementation of color image processing. The research showed the conversions of various models to speed up the image processing with least time delays.

Krishna, Neelima, Harshali, Rao (2018) studied the image classification using convolutional neural networks (CNN). They worked on four different images and observed that the images are classified correctly.

Some of the studies that used image processing on book cover to predict the genre like this study are as following.

Chiang, Ge and Wu (2015) used book title and cover image to predict genre. They used word2vec to find the probability of the title belonging to any genre category. They used several algorithms for image classification.

Iwana, Rizvi, Ahmed, Dengel and Uchida (2017) showed that a CNN can extract features and learn underlying design rules to define genre. They did not use any text data but stated as a future work that genre classification can also be done using textual features alongside the cover images.

Buczkowski, Sobkowicz and Kozlowski (2018) tried to make guesses about a book based on cover image. They used CNN. They saw that prediction based on textual description, which they studied previously on, are more accurate than cover image. However they achieved promising results.

3. Methodology

For this study Python is used on Google Colaboratory. Libraries that are used are Scikit-Learn, Pandas, Numpy, Keras, Matplotlib, Nltk, Imbalanced-Learn, Gensim, Scipy, Seaborn, Difflib, Cv2.

3.1. Prediction of Rating

3.1.1. Dataset and Feature Extraction

Properties of the dataset are summarized in Table 1.

DATASET	
# Samples	54.301
# Features	12
Numerical Features	# Ratings, # Reviews, # Pages
Text Features	Author, Title, Genre, Description, ISBN, Format, Edition, Image Url
Label	Book Rating

Table 1. Dataset Structure

"ISBN", "Format" and "Edition" features are not used in this study. "Image Url" is not used in the 1st part of the study.

After removing duplicated records, correlations between numerical features are calculated and due to the high correlation between "Number of Ratings" and "Number of Reviews", "Number of Ratings" feature is dropped. Afterwards, outliers are removed. As an example, Figure 1 exhibits the outliers having number of reviews more than 2.000.

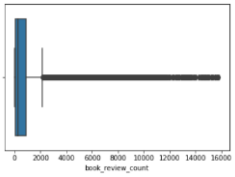

Fig. 1. Boxplot of (#Reviews) Feature

Samples with non-English book descriptions are removed. Final dataset has size of (38602,6).

There are 4 features in text format to be transformed: Author, Title, Genre and Description

Author is an important factor of success of the books. This feature is converted to ordinal numerical value by giving the authors points according to their overall success. Top 200 authors are specified from "www.ranker.com" which is a platform that people vote for various categories. Authors in top 200 list are given *5 points* whereas authors that are not in top 200 list are given *4 points*. Some books are written by more than one author. For them, the author with the best point is regarded.

Description, Title and Genre are converted to new features by applying topic modelling by LDA method.

The idea of topic modelling is that a document can be represented as a mixture of topics. Document processing takes places as following:

1. For each document, a topic is picked from its distribution over topics.
2. A word is sampled from the distribution over the words associated with the chosen topic.
3. The process is repeated for all the words in the document.

After the tokenization, stop word removal, stemming, lemmatization processes Description, Title and Genre features are applied LDA method. While applying LDA method, number of topics should be chosen to find the most qualified topic distribution. The number is chosen according to coherence scores

corresponding to the number of topics of the model. An example of Coherence plot is shown in Figure 2. According to this, number of topics to be used in LDA model is chosen 8 for title feature.

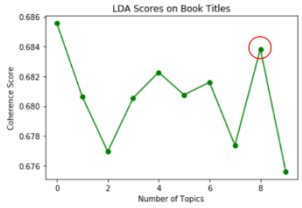

Fig. 2. Coherence plot of "Title" Feature

"Genre" feature, in fact, could be applied label encoding, or partitioned to groups by just looking over, if the number of genres were reasonable. However there are 2.634 genres due to the fact that a book is defined as composition of unique genres, e.g. a book is both classics, crime and horror. Figure 3 summarizes the situation.

	Genre	Number
1	Classics Christian Fiction	20
2	Childrens Picture Books	97
3	Classics Gothic Fiction	36
4	Parenting Autobiography Memoir	221
5	Fiction Sports and Games Sports	11
6	Thriller Fiction Suspense	28
7	Fantasy Paranormal Romance	317
.		
.		
.		
2634	Crime Politics	18

Fig. 3. Display of genre descriptions

Topic modelling created 8 topics for "Book Title", 5 topics for "Genre" and 5 topics for "Book Description". As a result (38602,21) sized dataset is ready for model fitting.

After standardization, data is randomly split with 75:25 training/test ratio. Most important features are acquired by Random Forest Classifier. Figure 4 shows the most important features chart.

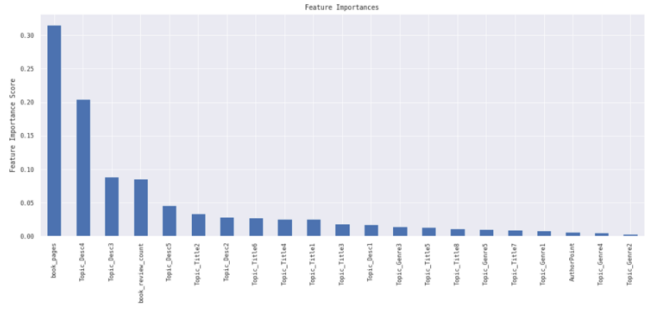

Fig. 4. Importance of features

As it can be seen, topics of descriptions are much more correlated with the rating compared to topics of titles and genres. This can be explained by the fact that if the text is longer, word co-occurrence of highly related words is more probable than that of short text. This means ensuring a more qualified topic modelling.

Several regression algorithms are applied on original features and feature subsets, like most important 5 features, 3 features, etc.

3.1.2.Results

Linear Regression, Xgboost, Random Forest, Gradient Boost, Lasso, Elastic, Neural Network algorithms are run on the dataset. Below, Table 2 gives the RMSE values comparison, each RMSE corresponds to the best result of the regressor, that is run on feature subsets.

	BEST RMSE
LR	0.266
XGBOOST	0.280
RF	0.264
GR.BOOST	0.249
LASSO	0.266
RIDGE	0.249
ELASTIC	0.253
SVR	0.271
NN	0.286

Table 2. RMSE values of Regressors

Maharjan et al. got a MSE value of 0.125 in their study which they used book content as feature. This study achieved to get a MSE approximately 0.07 on the average of the models.

Among the models the one which showed best performance is Neural Network model when it is examined in terms of loss curve. (Figure 4)

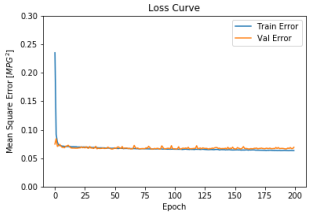

Fig. 4. Loss Curve of a NN model

3.2. Prediction of Genre

3.2.1.Dataset and Feature Extraction

For the second part of this study, “Image url” feature is to be processed. After removal of the disabled image url s from which the data can not be read, the number of the samples is 34.956.

Using Cv2 Python library, images are decoded to color data. Cover images are transformed to vectors which have RGB values as arithmetical averages of the pixels belonging to the image. The dimension of the vectors are (1x3), data corresponding to Red, Green and Blue color values respectively.

For title, 8 topic modelling features that are already extracted, are used. That is, the dimension of the final dataset is (34956,11)

This problem is not treated as multi-label classification. There are 2.634 types of genre definition as shown in Figure 3, so even if clustering is applied, it is difficult to decide on the number of clusters. In this study, the question is “Is the genre “Horror” or not?”. Genre is the binary label to be predicted. Classes are “Horror” labeled as 1, “Other” as 0. As a result, a row of the dataset to be worked on is shown in Table 3.

RED	GREEN	BLUE	TOPIC 1	TOPIC2	. . . . .	. . . . .	TOPIC 8	HORROR OR NOT
30.47	72.64	102.54	0.562	0.062	. . . . .	. . . . .	0.062	1

Table 3. Display of dataset for genre detection problem

Data is imbalanced with 7:93 class ratio (Horror:Other). Oversampling is applied. Data is randomly split with 75:25 training/test ratio, while maintaining the distribution of “Horror” and “Other” classes. Most important features sorted can be seen in Figure 5.

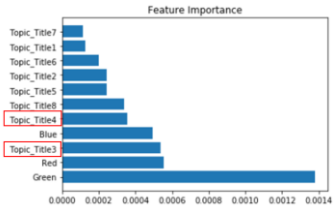

Fig. 5. Importance of features

Topic 3 and 4 are among the top 5 most important features. If we examine them in terms of most relevant words, 10 words for each are listed below. It can be observed that the words have meanings that reminds horror books.

Topic 3 : death, house, man, fire, god, blood, daughter, heart, find, end
Topic 4 : life, night, dream, lady, star, king, lie, big, white, good

3.2.1. Results

Support vector classifier and Logistic Regression classifier are applied on both all original features and subsets of them, like top 2 features, 3 features, etc. Area under curve (AUC) plot of the best model can be seen in Figure 6.

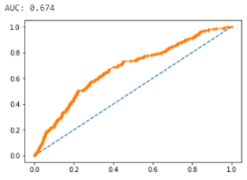

Fig. 6. AUC plot

Relevant classification report and confusion matrix are given below.

	Precision	*Recall*	*F1*	*Support*
0	0.95	0.56	0.71	3148
1	0.10	0.64	0.17	236
Accuracy			0.57	3384
Macro Avg.	0.53	0.60	0.44	3384
Weighted Avg.	0.89	0.57	0.67	3384

Table 4. Classification Report

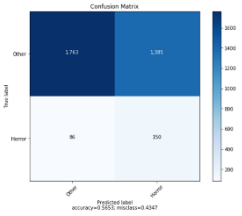

Fig. 7. Confusion Matrix

CONCLUSIONS

This study was an effort to see how much topic modelling in text can be useful for book rating prediction and how much book image and title can represent the genre. According to the results, it can be said that topic modelling did not work well to give an idea about the importance of book description, title and genre for rating prediction. There are other topic modelling approaches like LDA. They can be tried to see the difference. Adding new features, such as "Number of genres per author" can be tried to see if it enhances the model. Using deep learning techniques for text feature extraction can be utilized.

Taking color distribution of the cover of the book together with book title as inputs did not give satisfying results. Nevertheless, the topics 3 and 4 – words reminding someone horror related notions - taking the 3th and 5th places in most important features list gave clue that topic modelling should not be put aside for future works. The fact that "Green" and "Red" colors are the top 2 most important features can be inferred as even simple RGB representation of the image is useful for genre prediction. Although not yet to be sure, mixture of red and green colors is brown and we often encounter horror books covers with dominating brown color. This is just an intuitive inference. For the future work pixel neighborhood can be taken into consideration for a better modelling. Larger dataset will more likely achieve better results with the methodology of this study and other algorithms that are not tried in this study.

REFERENCES

D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, January 2003.

Maharjan, S., Arevalo, J., Montes, M., González, F. A., & Solorio, T. 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 (Vol. 1, pp. 1217-1227), 2017

D. Blei and J. Lafferty. Topic models. Text Mining: Theory and Applications, 2009

D. Alikaniotis, H. Yannakoudakis, M. Rei. Automatic Text Scoring Using Neural Networks. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016

Z. Tong, H. Zhang. A Text Mining Research Based on LDA Topic Modelling. Conference: The Sixth International Conference on Computer Science, Engineering and Information Technology, 2016

H. M. Wallach. Topic Modeling: Beyond Bag-of-Words. Proceedings of the 23 rd International Conference on Machine Learning, Pittsburgh, PA, 2006

X. Wang, A. McCallum, X. Wei. Topical N-grams: Phrase and Topic Discovery, with an Application to Information Retrieval. Seventh IEEE International Conference on Data Mining (ICDM 2007)

A. Arseniy and L. Natalia. Bigram Anchor Words Topic Model. International Conference on Analysis of Images, Social Networks and Texts, 2016

E. Laxmi Lydia, P.Govindaswamy, SK.Lakshmanaprabu, D.Ramya. Document Clustering Based On Text Mining K-Means Algorithm Using Euclidean Distance Similarity. Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 02-Special Issue, 2018

A. R. Afonso, C. G. Duque. Automated Text Clustering of Newspaper and Scientific Texts in Brazilian Portuguese: Analysis and Comparison of Methods. JISTEM - Journal of Information Systems and Technology Management, Vol. 11, No. 2, pp. 415-436, 2014

A. Hadifar, L. Sterckx, T. Demeester, C. Develder. A Self-Training Approach for Short Text Clustering. Proceedings of the 4th Workshop on Representation Learning for NLP, 2019

R. J. Mooney, P. N. Bennett, L. Roy. Book Recommending Using Text Categorization with Extracted Information. AAAI-98 Workshop on Recommender Systems,pp.49-54 and pp.70-74, Madison, WI, July 1998

B. K. Iwana, S. T. R. Rizvi, S. Ahmed, A. Dengel, S. Uchida. Judging a Book by its Cover. Kyushu University Education Reform Symposium, Fukuoka, Japan, 2018

P. Buczkowski, A. Sobkowicz, M. Kozlowski. Deep Learning Approaches towards Book Covers Classification. Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018), pages 309-316, 2018

J. Zujovic, L. Gandy, S. Friedman, B. Pardo, T. N. Pappas. Classifying Paintings by Artistic Genre: An Analysis of Features & Classifiers. IEEE International workshop on. IEEE, 2009.

M Manoj Krishna, M Neelima, M Harshali, M Venu Gopala Rao. Image classification using Deep learning. International Journal of Engineering & Technology, 7 (2.7) (2018) 614-617, 2018

A. A. Kulkarni, R. D. Kurundkar, S. V. Khare, S. Savant, , P. Chintal. A Review on Color Image Processing. International Journal of Computer Science and Mobile Applications, Vol.3 Issue. 11, November- 2015, pg. 5-10 IS

https://www.kaggle.com/data13/predict-book-rating-with-linear-regression

https://www.ranker.com/list/best-writers-of-all-time/ranker-books