International Journal of Innovative Research in Engineering
Volume 3, Issue 3 (May-June 2022), PP: 250-253.
www.theijire.com
ISSN No: 2582-8746

A Survey on Book Genre Prediction Methods from Summary

Aditya Krishnan[1], Alwin Antony[2] ,Durga Jayachandran[3] ,Shoba T[4],
[1,2,3,4]*Computer Science, Adi Shankara Institute of Engineering and Technology, Kerala, India.*

How to cite this paper:
Aditya Krishnan[1], Alwin Antony[2] ,Durga Jayachandran[3] ,Shoba T[4], "A Survey on Book Genre Prediction Methods from Summary", IJIRE-V3I03-250-253.

Abstract: *The importance of a book summary is that it gives readers a first impression of the novel. The classification of a book's genre is dependent on its summary, and it is critical in modern systems. Given that comprehensive digitization of books is a prohibitively expensive undertaking. At the same time, 4it is a difficult task because there are many different book genres, and the summary will vary depending on the theme, textual information, and other factors. Even for books in the same genre, the summary may vary depending on many external factors such as country, culture, target reader populations, and so on .In today's generation, text classification is quite crucial. Because the number of digital users is growing every day, its demand is growing as well. As a result, machine learning algorithms are employed to classify specific text input, resulting in more accurate predictions. The genre of a book can be predicted by generating a data set with correct organization and data. The book title and summary will be included in the datasets. The major goal was to use 15machine learning techniques to classify a book by its genre*

Key Word: *Text Classification, Machine Learning, RNN, LSTM, Multi label Classification*

I. INTRODUCTION

In human history, books have been the most essential medium for recording information and knowledge. Books are divided into 2categories based on their covers, contents, languages, and other factors. The job of categorizing books by genre using the information provided in the summary is the topic of this paper. The initial impression a reader gets from a book summary is that it conveys significant 4information about the book's substance. Data pre-processing, Text cleaning, Feature Selection, Training Model, Assigning Classifiers, and Output are the processes that most text classification can be broken down into. The most common issue in libraries nowadays is 6the classification of books by genre. Many books are not classified by genre, which makes it difficult for librarians and readers to classify them. Predictions of genres are created 1based on the book titles and summaries to classify the genre of these novels. The idea is to use the summary to anticipate the book's genre. A dataset with title, author, characters, and narrative will be used. This data will be used to categorize and predict the book's class. The goal was to create a properly genre-categorized book collection that would allow customers to quickly find books that they needs. Also, it may 1be used to organize books in large bookstores and libraries according to their needs.

II.MATERIAL AND METHODS

SUPERVISED LEARNING METHODS

K-Nearest Neighbor (K-NN):

It is one of the simplest algorithms based on supervised learning technique. The 8K-NN algorithm assumes the similarities between the new case/data and the available cases and places the new case in the category that is more similar to the available categories. It stores all the data that are available and classifies a new data point based on similarities. It 1can also be used for regression .K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data .So that the new data can be rapidly grouped into a well-defined group using the K-NN algorithm. It uses the Euclidean Distance Formula and the Manhattan Formula to determine similarity. The distance between data points is determined using the K-NN algorithm.

Decision tree:

Decision tree is the most powerful and popular tool for classification and prediction. Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label .A non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

Logistic Regression:

Logistic regression is a supervised classification algorithm and a binary classifier .This regression is generally used to separate data into two classes. Multinomial logistic regression can be used to classify data into three or more classes. Logistic Regression is a model that has been formed with the use of Logistic Function. The Sigmoid Function is another name for this Logistic function.

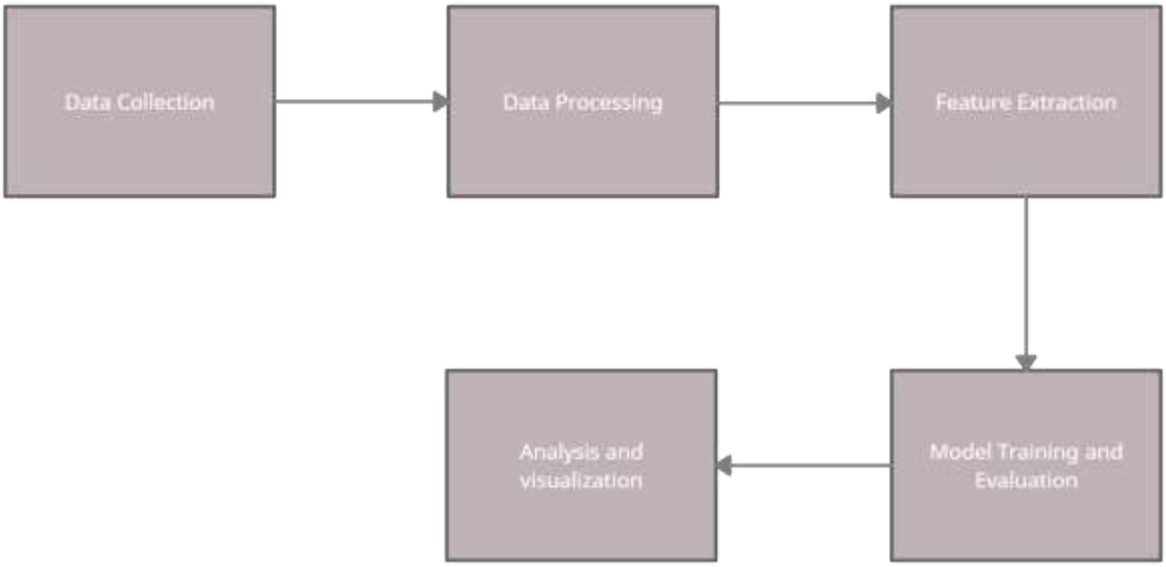
DEEP LEARNING METHODS

RNN-LSTM :

Recurrent Neural Networks are a more general form of neural network that can remember past patterns. RNN is inherently recurrent in that it performs the same function for each data input, while the output of the current input depends on the last computation. After producing the output, it is copied and sent back into the recurrent network. This allows the network to learn from the past and make predictions for the future. For making a decision, it considers the current input and how that input has affected the output it has learned from the previous input. Long-term memory (LSTM) networks are a modified version of recurrent neural networks that make it easier to recall past material. The problem of RNN's vanishing gradient is handled here. Given time lags of uncertain duration, LSTM is well suited to identify, analyse, and predict time series. Using back-propagation, it trains the model.

METHODOLOGY

PLAN :



DATA SET :The data set we’re using consists of book descriptions and genre classification is scraped from CMU Book Summary Dataset . This dataset contains plot summaries for 16,559 books extracted from Wikipedia, along with aligned metadata from Freebase, including book author, title, and genre. Each book has the following metadata: Wikipedia ID, Freebase ID, Book title, Book author, Publication date, Genres, Summary. Only Summary and Genre is taken for modeling.

DATA PREPROCESSING: The summary is then processed by eliminating all punctuation, uncasing all letters, and tokenizing the words. Note that the books themselves are almost too long to be feasible 1for training the classification models, with most of them having tens of 6thousands of words, so we decide to classify the summaries of these books, which are relatively short and straightforward .This is step take longest time . Observing the data, common problems we face are Missing values ,Different languages involved ,non-Ascii characters, Invalid descriptions ,Missing spaces in the description. Main steps include:

- Remove descriptions with invalid format
- Get English descriptions only for simplicity
- Clean The Text so that non-Ascii characters
- punctuations and other symbols are removed
- All text will 1be converted to lowercase
- The most common words also known as stop words are removed
- Lemmatization and Stemming is performed.

FEATURES EXTRACTION: TfidfVecfunction from sklearn library is used to extract features from abstract and assigning weights to the feature values Not only TfidfVectorizer, any feature extraction techniques can be used for extraction of feature values from data like 15bag of words, etc. tf–idf or TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus .We will set a description limit. Then we denote an unique number 2to each of the genres using label encoder.

METRICS: Several classification measures, such as accuracy, precision, recall, and F-1 score, are used to evaluate model performance. The proportion of correctly categorized samples among all the samples in the dataset is a typical statistic for classification problems. However, because each book has a different amount of labels, it's incredibly difficult for the model to match the goal output completely. To loosen this constraint, we first compute the number of samples the classifier has made in terms of true positive predictions (T P) and true negative predictions (T N) for each label class c. (T N). The weighted average of correctness across all label classes is then used as our accuracy metric. It can be written more properly as

accuracy = ∑c∈Cwc · 1NN∑i=1 1[(c ∈ f (xi) ∧ c ∈ ˆf (xi)) ∨ (c /∈ f (xi) ∧ c /∈ ˆf (xi))],

where f signifies the learnt classifier and f denotes the ground truth function The multi-label output of the ground truth function for observation xi would be represented by f (xi), while the multi-label output of the learnt classifier for the same observation would be represented by f (xi). The label class frequency is used to calculate the weights. Because labels with a higher frequency are more important than labels with fewer observations, labels with more observations are given higher weights. We normalize the weights by dividing each label frequency by the sum of all frequencies to verify that they all add up to one.

When computing other metrics for the multi-label output, such as precision, recall, and F-1 score, we use the same reasoning. We compute the metric for each label class and average it out using the (normalized) number of samples for each label. As a result, all of our assessment criteria have been able to account for label imbalance in order to more accurately reflect model performance.

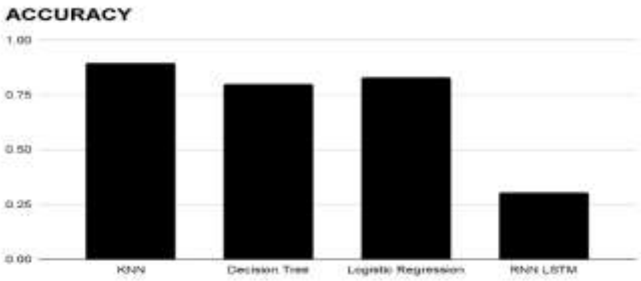
EVALUATION: Dataset was split into a training set, a validation set and 6a test set, We split our dataset into a training set, a validation set and a test set, with an approximate 80- 20% split ratio. In our setup, a preprocessed text was provided to a model and was tasked to output a multi-label prediction Y ∈ 2{0,1,··· ,C−1} where 2the total number of classes to consider is . We used this setup for all models we were to explore:, logistic regression, decision tree,, 6as well as the RNN LSTM -based deep network. We compared their performance using our custom metrics, and then looked into the generalization capabilities of our models by inspecting 2the distribution of predicted labels**.**

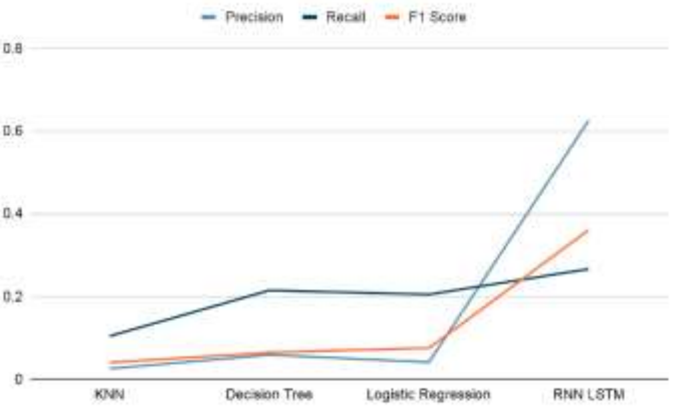
$$f(x) = \left\{ \underset{0\le c<C}{\operatorname{argmax}} \sum_{x_i \in X_{train}} \mathbb{1}[c \in f(x_i)] \right\}$$

III.RESULT

Graph 1 compares the performance of the various models using accuracy. KNN has the best accuracy .Considering the other metrics on the graph 2, the Decision tree has performed better. Could be fairly surprising at first sight to see that the N -nearest neighbour baseline is on top in terms of accuracy among all learning models. But this happens to reflect the exact drawback of this metric—a large number of true negatives predicted by this baseline can inflate its accuracy score to even be comparable to more complicated models. Therefore, it would make more sense to look at theF1 score, which is a harmonic average over precision and recall. From F1 scores, we observe that RNN LSTM outperforms the other methods. Probabilistic models have more F1 score This might be caused by the fact that probabilistic models are operating in the real fields which give more model capacity, while decision-tree based methods are mostly based on counting and statistics.

Graph 1





Graph 2

IV. DISCUSSION

As the amount of data grows exponentially in the modern day, the need for text data classification and categorization grows as well. Machine learning techniques might be quite useful in resolving this problem. It has also been observed in venues such as libraries, bookstores, and eBook sites where books are not classified according to their category. In this survey paper, we explored how to predict a book's genre given its summary. We implemented various machine learning models to facilitate this process, including N -Nearest Neighbour, Logistic Regression, Decision Tree and Deep Learning model. KNN has the best accuracy. We TF-IDF and IGF were used to make contributions to feature engineering (Inverse Genre Frequency) To reduce training and inference time and improve model performance, features and a paragraph selection heuristic were included performance. We were able to accomplish text categorization on extremely lengthy texts this way. To Deal with problems in multi-label classification, we define a probability threshold and For multi-genre inference, paragraph-majority votes were used. We created a unique assessment system. parameters in order to more accurately understand our findings.

V. CONCLUSION

KNN clearly had highest accuracy compared to other machine learning methods taken for this survey.It is also observed that Deep learning method had outperformed other methods in metrics such as precision,recall and F1 Score

References

[1]Avinash Navlani,KNN Classification Tutorial using Scikit-learn,Retrieved from https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn?utm_source=adwords_ppc&utm_medium=cpc&utm_campaignid=1455363063&utm_adgroupid=65083631748&utm_device=c&utm_keyword=&utm_matchtype=&utm_network=g&utm_adpostion=&utm_creative=332602034364&utm_targetid=dsa-429603003980&utm_loc_interest_ms=&utm_loc_physical_ms=9040212&gclid=CjwKCAjwur-SBhB6EiwA5sKtjmE-wBv8_PGBRB2YtTksuzTnxFwVdwzOG5WkkRAi7g9qB4BKcS2TfhoCbVIQAvD_BwE

[2] JOUR, Shiroya, Parilkumar,Vaghasiya, Darshan,Soni, Meet,Panchal, Brijeshkumar, Book Genre Categorization Using Machine Learning Algorithms (K-Nearest Neighbor, Support Vector Machine and Logistic Regression) using Customized Dataset, International Journal of Computer Science and Mobile Computing,2021.

[3]Sicong, Liu Zihan Huang ,Yikang Li, Zhanlin Sun ,Jiahao Wu, Hongyi Zhang,DeepGenre: Deep Neural Networks for Genre Classification in Literary Works ,Language Technologies Institute Carnegie Mellon University

[4]Varshit Battu,, Vishal Batchu, Rama Rohit Reddy,, Murali Krishna Reddy,Radhika Mamidi,Predicting the Genre and Rating of a Movie Based on its Synopsis,International Institute of Information Technology Hyderabad.

[5]Aidan Finn , Nicholas Kushmerick,Learning to classify documents according to genre,Smart Media Institute, Department of Computer Science, University College Dublin

[6]Dr. S. Rama Krishna,Dr. S. V. Vasantha,K. Mani Deep,Survey on Fake News Detection using Machine learning Algorithms,International Journal of Engineering Research & Technology (IJERT).

[7]Meenakshi,Machine Learning Algorithms and their Real-life Applications: A Survey,International Conference on Innovative Computing and Communication (ICICC 2020).

[8] Saloni Gupta(22 Jun, 2021),Decision Tree, Retrieved from https://www.geeksforgeeks.org/decision-tree/

[9] https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

[10] http://www.cs.cmu.edu/~dbamman/booksummaries.html

[11] Prateek Joshi, Predicting Movie Genres using NLP-An Awesome Introduction to Multi-label Classification,available:-https://www.analyticsvidhya.com/blog/2019/04/predicting-movie-genres-nlp-multi-label-classification/

[12] Akshay Bhatia, Book-Genre-Classification, available at:https://github.com/akshaybhatia10/Book-Genre-Classification

[13]David Bamman and Noah Smith (2013), "New Alignment Methods for Discriminative Book Summarization," [ArXiv] booksummaries.tar.gz [17M]

[14] Aditi Mittal(Oct 12, 2019), Understanding RNN and LSTM, Retrieved from Understanding RNN and LSTM. What is Neural Network? | by Aditi Mittal | Medium