Prediction of Genre from Book Titles

Siddharth Shenoy
PES2201800499
Computer Science Dept
PES University

Apoorva Choudhary
PES2201800102
Computer Science Dept
PES University

Reshma P Roy
PES2201800039
Computer Science Dept
PES University

***Abstract* — This paper compares various methods for producing a book's genre based on its title. Genre classification is a form of text classification. It is a means of managing information. As text databases becomes larger, genre becomes increasingly important. Three learning models were tested throughout the experiment namely Bag of Words, Logical Regression and Naive Bayes. Each algorithm was fine tuned for attaining the best parameter values. Basic analysis of the dataset was also done to find trends and patterns when it comes to a book's title with respect to its other features. The proposed system is implemented in python programming.**

Keywords — genre prediction, regression, bag of words, naive bayes, book title, classification system

I. Introduction

Genre classification is the process of grouping objects together based on defined similarities such as subject, format, style, or purpose. Classifying books into genres helps us catch a glimpse of what the book is about, tells us about the contents of the book and helps us decide if we will like it or not. Genre classification is already very well established in some fields of arts like music. It is not that well defined when it comes to text. Document genre, as with music, pertains to style and form.

Genre classification is necessary to solve many of the problems faced by computational linguists. Parsing accuracy could be increased by taking genre to account. In information retrieval, genre classification could enable users to sort search results according to their interests. People hunting for books usually select books by genre. Even scholarly articles are divided into categories such as history, machine learning, deep learning etc.

Recently, there has been an increasing interest in automated genre classification. This is an essential step in managing documents according to organisational activities. Different kinds of features have been employed as input to classification models which have been shown to achieve high accuracy in classification scenarios under controlled environments. Automated book classification is generally defined as content-based assignment genres to books. Classification of literary works is significantly different from normal text classification. One big reason for this is length, as books are generally much longer than most other text mediums. Parsing through the entire book just to classify it can be a tedious task and hence we try to classify it just based on its title as titles are usually indicative of the book's content. Book titles are an important part of a book's presentation. The aim of this project is to analyze these titles and choose an optimal model for classifying them. The problem with book classification is that there is no correct set of rules for classifying a book into a particular genre. Genre definitions can differ based on society, region and person. So it is important that books be given a degree of relativity to a given genre.

The models test traditional methods of classification such as bag of words, logistic regression and naive bayes. Our dataset contains details of books along with authors and ratings. Each book has a list of genres that have been voted upon by the readers.

II. RELATED WORK

This project builds on traditional methods of text classification and sentiment analysis. The first linguistic research on genre that uses quantitative methods is that of Biber (1986; 1988; 1992; 1995), which draws on work on stylistic analysis, readability indexing, and differences between spoken and written language. Biber ranks genres along several textual

"dimensions", which are constructed by applying factor analysis to a set of linguistic syntactic and lexical features. Those dimensions are then characterized in terms such as "informative vs. involved" or "narrative vs. non-narrative." Factors are not used for genre classification (the values of a text on the various dimensions are often not informative with respect to genre). Rather, factors are used to validate hypotheses about the functions of various linguistic features.

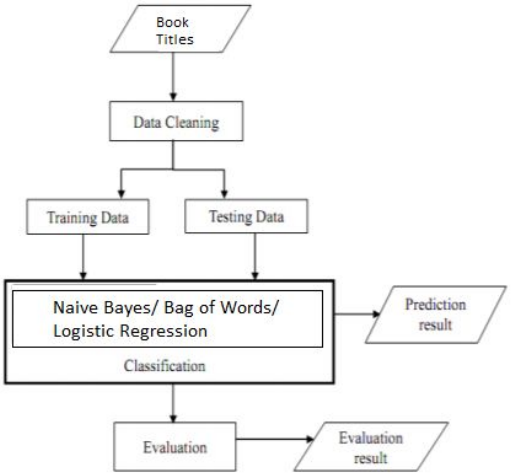
Other research has dealt with using book contents and book summaries for classification. Text classification includes techniques such as Naive Bayes, Tf-idf, latent semantic indexing, support vector machines, decision trees, and natural language processing. The models we have chosen which are Naive Bayes, Bag of Words and regression are proven to give good results. Many of the recent work done on text classification deals with algorithmic comparison using different machines like recurrent neural networks, long short-term memory, convolution neural networks and more. Most projects deal with the problem of developing approaches which are capable of working with extremely large and complex literature documents. Word similarity was another explored approach by other people. Book summaries were examined to find word similarity. Bag of words along with score comparison and percentage comparison was used to identify the genre of books. Other features to be noted are that due to possible imbalance in data sets, MLP tended to perform poorly, while Decision Trees and Ensembles of Random forests performed much better

We try to minimize computing power by only analysing the titles of books which are good indicators on what a book is about and its contents. Since books are written creatively, a tree structure based algorithm is the best way of predicting genre. Therefore Bag of Words is used. Bag of words can be made as complex as needed for a given problem statement. Naive Bayes is a test classifier based on the Bayes theorem, which helps us compute the conditional probabilities of occurrence of two events based on the probabilities of the occurrence of each individual event, encoding those probabilities is extremely useful. Logistic regression was chosen as a basic numerical method. Other papers indicated that logical regression gave better results than linear dicrimination and linear regression. For binary classification, the application of logistic regression was straightforward.

III. PROPOSED SOLUTION

This project explores three traditional models - Bag of Words, Naive Bayes and Logistic Regression.

The implementation block of Automated Book Classification is given below



Bag Of Words

A bag-of-words model, or BoW for short, is a way of extracting features from the text for use in modelling, such as with machine learning algorithms. It is a way of extracting features from the text for use in machine learning algorithms. One of the biggest problems with text is that it is disordered and unstructured, and machine learning algorithms prefer structured, well defined fixed-length inputs and by using the Bag-of-Words technique we can convert variable-length texts into a fixed-length vector. Also, at a much granular level, the machine learning models work with numerical data rather than textual data. So to be more specific, by using the bag-of-words (BoW) technique, we convert a text into its equivalent vector of numbers.

In this approach, we use the tokenized words for each observation and find out the frequency of each token. We treat each sentence as a separate document and we make a list of all words from all the four documents excluding the punctuation. The next step is to create vectors. Vectors convert text that can be used by the machine learning algorithm. We check the frequency of words from the 10 unique words. In this approach, each word or token is called a "gram". Creating a vocabulary of two-word pairs is called a bigram model.

We removed a few rows and a few columns as part of pre-processing as they were not a part of the training or testing algorithm. As there were a huge amount of redundant genres which had to be generalised to standard genres to get solid numbers of accuracy. We built the model thrice to check the maximum accuracy we could get with 10 epochs. Each the accuracy came around 0.49-0.51 which is good as our approach is limited to predicting genres from just the title. Logistic

Regression was used to explore a basic numerical method of classification.

Naive Bayes

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.There are three types of Naive Bayes Model, Gaussian, Multinomial, Bernoulli.

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.P(A) is Prior Probability: Probability of hypothesis before observing the evidence.P(B) is Marginal Probability: Probability of Evidence.The fundamental Naïve Bayes assumption is that each feature makes an independent and equal contribution to the outcome. This assumption makes the Bayes algorithm naive. Given, n different attribute values, the likelihood now can be written as

$$P(X_1 \ldots X_n|Y) = \prod_{i=1}^{n} P(X_i|Y)$$

Here, X represents the attributes or features, and Y is the response variable. Now, P(X|Y) becomes equal to the products of, probability distribution of each attribute X given Y. After preprocessing the data, we implement the model and fit it with the data accordingly.

Preprocessing involved removing rows where the titles and genres were missing. Only the highest rated genre for each book was considered. The number of unique genres were around 100 therefore they were manually reduced by regrouping similar genres together. Before applying the model, unnecessary words, characters or numbers were removed from the “title” column. This provides the model to train on only those words that are important in determining the genre of a book. The model is then implemented on the final dataset.

Logistic Regression

LR is a statistical technique for modeling a binary response variable by a linear combination of one or more predictor variables, using a logit link function: $g(\pi) = \log(\pi/(1 - \pi))$ and modeling variance with a binomial random variable, i.e., the dependent variable $\log(\pi/(1 - \pi))$ is modeled as a linear combination of the independent variables. The model has the form $g(\pi) = x_i\beta$ where π is the estimated response probability (in our case the probability of a particular facet value), x_i is the feature vector for text i, and β is the weight vector which is estimated from the matrix of feature vectors. The optimal value of β is derived via maximum likelihood estimation.

For our LR model, the response variable genre is modelled against the predictor variable book title. Preprocessing involves dropping columns that weren’t necessary and resetting indexes. As the dataset had genres that were voted by the users, there was a wide variety of genres which had to be generalized into standard genres. Even after standardization, the genres which had less than 25 books under it were dropped as to not lead to loss in accuracy. The genres were then put through a label encoder. The titles were processed to remove stop words which were taken from the nltk.corpus library. The titles were then vectorized. The logistic regression model uses the SAG(stochastic average gradient) method as it works faster for large datasets. It achieves a faster convergence rate by incorporating the memory of previous gradient values. The test-train split is twenty percent. There are around 50 genres after the pre-processing and dropping of irrelevant rows from 183 genres. The final accuracy is around 46 to 52 percent.

IV. RESULTS

As we all know, the title of a book is a very creative way of telling what’s inside the book but many a time the creativity leads to some kind of vagueness. As a result a tree architecture would be ideal to predict the genre of a book. This is why the Bag of Words model was implemented. Moreover, the genres present or mentioned in the data were so spread out that choosing a genre and relating it with the word would provide us so less data that the accuracy of the model would be negligible which forced us to generalise the genres.

Basic analysis of our dataset revealed popular genres based on their ratings. Graphs were also used to understand the correlation between the average rating and the rating which affects the genre the most. A correlation matrix was built to examine the relationships between features. The mean, median and sum of genre votes along with the distribution of books tagged with an individual genre was also visualized. We thus get a general idea of popular genres and can also identify outliers which need to be removed.

Further analysis was done to see if the genre was consistent. For example, all the books in the same series were analysed to see if they fell under the same genre which ended up to be true. Authors also wrote books mostly falling in the same genres.

The final results of our report were not the best as were faced with the limitation of predicting the genre just by its title and without using deep learning techniques like neural networks. The models just pass the 50 percent mark with Naive Bayes performing the worst. Another step back in our dataset was the inconsistency in the assigning genres. As the genres for the test were assigned by readers there were no standard genres allotted which ended up in a varied genre dataset with many outliers. The dataset was cleaned to a certain extent by trying to merge similar genres together and dropping books that belong to genres with very low frequency.

The main disadvantages of the models used are:

In the bag of words approach, we see the curse of dimensionality as the total dimension is the vocabulary size. This leads to over-fitting. Bag of words also doesn't consider the semantic relation between words. In text classification, the neighbourhood words usually play an important role in predicting the target word. When it comes to Logistic Regression, it doesn't perform well when the feature space is too large. Naive Bayes assumes that all features have the same statistical relevance which is not the case.

CONCLUSION

Contribution of team members:

Siddharth Shenoy implemented the Bag of Words Model. Apoorva Choudhary implemented the Naive Bayes model. Reshma P Roy implemented the Logistic Regression model. Preprocessing and basic analysis was done in collaboration with each other.

REFERENCES

[1] Ozsarfati, E., Sahin, E., Saul, C. J., & Yilmaz, A. (2019, February). Book Genre Classification Based on Titles with Comparative Machine Learning Algorithms. In 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS) (pp. 14-20). IEEE.(Ozsarfati et al., 2019)

[2] Worsham, J., & Kalita, J. (2018, August). Genre identification and the compositional effect of genre in literature. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 1963-1973).

[3] Kyaw, T., Htun, Z. M., & Swe, K. T.(June, 2020) CLASSIFICATION SYSTEM USING HYBRID COMPARISON METHODS

[4] Kessler, B., Nunberg, G., & Schütze, H. (1997). Automatic detection of text genre. arXiv preprint cmp-lg/9707002..

[5] Ertugrul, A. M., & Karagoz, P. (2018, January). Movie Genre Classification from Plot Summaries Using Bidirectional LSTM. In Semantic Computing (ICSC), 2018 IEEE 12th International Conference on (pp. 248-251). IEEE..

[6] Emily Jordan.: Automated Genre Classification in Literature, (2014).