DATA ANALYTICS PROJECT
LITERATURE SURVEY

TEAM 7
D-LINK

PAPER 1:

Title: Book Genre Classification Based on Titles with Comparative Machine Learning Algorithms
Authors: Eran Ozsarfati, Can Jozef Saul, Egemen Sahin, Alper Yilmaz

Citation: 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.

This paper presents algorithmic comparisons for producing a book's genre based on it's title. Five different machine learning titles were tested to determine the most optimal and accurate method. They were: Recurrent Neural Networks (RNN), Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), BiDirectional LSTM (Bi-LSTM), Convolutional Neural Networks (CNN), and Naive Bayes.
The data was tokenized and normalized to create a custom dictionary. The resulting data was separated into words and stemming was applied to the data, reducing the vocabulary size without losing information. Then they were put through classification algorithms to get an output. Multinomial Naive Bayes implements multinomial distributions and thus implements the frequency of words in the dataset. CNN is good at detecting local patterns by creating feature maps. A nonlinear function (ReLU) is applied to make sure there are no negative values. The CNN model in this experiment uses three convolution layers and 3 max-pooling layers. RNN is useful in tasks such as word predictions where previous words are required to make a good prediction. RNNs can capture long term dependencies but in practice, they fail to do so because of the vanishing gradient problem that arises because common activation functions map the real number line onto a range of [0,1]. GRU and LSTM are able to solve this problem. GRU solves the vanishing gradient problem by using update and reset gates. These gates decide what information should be passed on or not. LSTM is more complex and advanced than RNN. It has four interactive gates with respective functionalities. Bi-Directional LSTM comprises one forward LSTM and one backward LSTM. Bi-LSTM concatenates the final hidden states and then applies the softmax function to get an output.
The results show a gradual improvement in the line of RNN variants, with LSTM being the top-ranked model. While GRU has more filtration operations and a longer memory than RNN, LSTMs additional binary classification increases the complexity giving it a 7.3% better accuracy than GRU. Although LSTM and Bi-LSTM conduct similar operations, LSTM yielded a higher accuracy due to the short length of book titles, unlike other multi-class classification

studies. The high-performance rate of the CNN shows its similarity to the LSTM architecture, with its ability to store information. LSTM slightly outperforms CNN as CNN is merely able to detect local patterns in a title, while LSTM takes the entire input to its memory, enabling it to find more vast scaled patterns in titles. The results from deep learning methods yielded the following order of increasing accuracies: RNN, GRU, CNN, Bi-LSTM and LSTM.

PAPER 2:

Title: Genre Identification and the Compositional Effect of Genre in Literature
Authors: Joseph Worsham, Jugal Kalita

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).

This paper addresses the problem of developing approaches which are capable of working with extremely large and complex literature documents to perform genre identification. It takes an approach where a literary genre is considered to be a writing style family where texts that contain similar themes are grouped together. It assigns a literary classification to a full-length book belonging to a corpus of literature. It also compares current deep learning models to traditional methods. The dataset used in this paper is the Gutenberg Dataset which is an extensive web catalogue containing over 56,000 ebooks. The project deals with large documents containing well over 200,000 words compared to traditional document modelling datasets which have around 100-1000 words. Traditional approaches such as bag-of-word(BOW) and bag-of-n gram models are naturally unable to capture information that persists across paragraphs and chapters. Themes that contribute to the assignment of a genre are not only based on the frequency of words but concepts presented over the entire work. Therefore, neural modelling approaches which learn fixed-length semantic models were considered. Networks like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks (LSTMs) or a combination thereof were explored. CNN becomes impractical as they grow increasingly larger with the input. Sequential models like RNNs and LSTMs are sensitive to long sequences. This work researches the effect of modelling smaller portions of text which are extracted from the greater work.

CNN-Kim was one of the first neural network implementations for text classification. It incorporated a CNN whose convolutions captured neighbourhoods of word embeddings. LSTM is a more complex network architecture that was proposed for document modelling and classification. It was designed to capture the natural structure of languages where word representations are composed into sentence representations which are in turn composed into document representations. The next model used in genre identification is Hierarchical Attention Network(HAN). The hierarchical nature of the HAN, plus the addition of the popular attention mechanisms make HAN a good fit for the Genre Identification problem. The encoding layer on HANis based on that of LSTM. The attention layers of the HAN are designed to learn which encoded elements from the previous layer.HAN has three components in it. The first is a simple MLP which would be trained to build annotation over each element. Second The measurement of the importance of the element is then computed using a softmax function. Finally, a weighted sum of the encoding layer using the importance measurement is produced as a representation for the element at the next hierarchical level

Along with the detailed deep learning models, they employed traditional machine learning models for classification in order to establish a performance baseline for the Genre Identification task. such as k-Nearest Neighbor, Random Forest and XGBoostare are all evaluated with the BOW input method.
KNN is different from the other models in that it simply employs a distance metric in order to identify which other documents in the dataset a new record is found to be close to.
The Random Forest model is made up of an asynchronous collection of small decision trees which are trained on random selections of features from the dataset.
XGBoost is a highly optimized, award-winning Gradient Boosting solution which is made up of a boosted set of sequential trees learned from the gradients of some differentiable loss function

Input:
As the data set is huge and terms in the book are directly related to the genre of the book diff strategies of input method is used
First 5,000:The first 5,000 input methodology both trains and evaluates on only the first 5,000 words in each record. In this case, the remainder of each book is discarded.
Last 5,000:The last 5,000 input methodology, like the first 5,000, trains and evaluates on only the last 5,000 terms found in the dataset.Again, there is no need for padding as every document is longer than 5,000 terms.
Random 5,000:: The random 5,000 method will randomly extract a 5,000 term window from a record on every training batch and evaluation batch.
All Chapters:This input methodology applies a simple pattern matching algorithm to each book in order to split the books up into chapters. In this case, entire books, both in the training set and testing set, will be processed.
Bag-of-words: BOW is a simple vector representation of an entire document where each dimension is representative of a unique word in the corpus. Thus, the input dimensionality is equivalent to the number of terms which are maintained when building the BOW representation.

Results :
Accuracy and the F1-Measure are used as a measure of a model’s performance. Each model had weaknesses when working with this dataset.
CNN-Kim had the most reliable performance of the deep learning models, scoring a total of 75% accuracy on the “All Chapters” input.

HAN model had a difficult time converging on this dataset, frequently being fooled by the more represented classes such as Adventure Stories and Science Fiction.
The LSTM, the lowest scoring neural network architecture, collapsed and was unable to learn relevant features over the long sequences.
XGBoost outperformed all models to have the highest accuracy at 84% and F1-Measure at 81%

This research has shown that across the corpus, there was a 6% gain in accuracy when training and evaluating on the first 5,000 words of each book over training and evaluating on the last 5,000. There was only an additional 2% increase in accuracy when training and evaluating on the random 5,000 words.

PAPER 3:

Title: Classification System Using Hybrid Comparison Methods
Authors: Thiri Kyaw, Zin Mar Htun, Khaing Thanda Swe

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

This paper aims at building an automated book classification which uses text-based comparison of book summaries to examine whether word similarity is a feasible method to identify the genre of the book. This system focuses on a bag of words approach, score comparison and percentage comparison method to classify genre. Bag of words approach is used to count the words in summary while score comparison and percentage comparison method is used to identify the genre of the book
Score comparison method classifies the books by giving them a total point score based on which words occur in the books. Firstly all the words from the book data are counted and then this complete list of words is examined and points are assigned based on their number of occurrences. It looks at each count and finds out the influence of the genres based on their points.
Percentage comparison method takes into account the size of the word collection for each genre. It does this by adding the total word counts for each genre and then by dividing the word counts by that total. The word counts for the unknown book summaries are calculated and once the percentages are tabulated, the percentages of similarity are added up and the genre with the highest total percentage of similarity is deemed as the apt genre.
The model is then tested after the dataset is divided into test and training dataset. The data is preprocessed before testing. Only required information like - title, genre, authors are kept. Other unwanted information is discarded.
Finally, evaluation is done in the terms of precision and recall. Cross-validation and sanity check are also tested. The final experimental results of the proposed system have over 80% for both precision and recall values.