National College of Ireland

Project Submission Sheet – 2020/2021

Student Name:	Omkar Ratnoji Tawade		
Student ID:	19232136		
Programme:	MSc. Data Analytics (MSCDAD_A)	**Year:**	2020-21
Module:	Data Mining and Machine Learning 2		
Lecturer:	Michael Bradford		
Submission Due Date:	24-05-2021		
Project Title:	TABA		
Word Count:	3818 words		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	Omkar Tawade
Date:	24-05-2021

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Case Study and Paper Review Analysis

Omkar Tawade

Data Analytics

National College of Ireland

Student Id: 19232136

- **Question 1** – Conduct a case study for Identification of legal precedents in case law .
- **Question 2** – Conduct a paper review of the paper "Wheat varieties identification based on a deep learning approach" published by Karim Laabassi et al. (2021).

1. Question 1 - Case Study

The first part of this report is based on a scenario where I have been hired as a data analyst to work on a machine learning project in the problem domain "Identification of legal precedents in case law". The case law is the collection of historical decision written by courts while deciding the case and these past decisions are known as case law. Case law is also known as judge made law. A case that establishes rule or principle which is used by court to decide the judgment of later cases with similar topic. We will be using the natural language processing method to identify the legal precedent in the case law. We will start our project by implementing the data cleaning step in which we will use few method of natural language processing to clean the data and make the data ready for modelling. Further we will implementing the feature selection, again here we will be using concept of corpus and word vectors which are the part of natural language processing. We will also discuss about the TF-IDF method and word embedding methods for feature extraction. In next step, we will discussing our modelling technique, we have decided to implement the convolutional neural network model, we will be discussing about the convolutional neural network in depth and also discuss why we used the convolutional neural network over other algorithms. After modelling technique we will discuss how we will gauge the performance of the model based on performance metric. In the end, we will discuss about the scalability issue which can avoided while executing the project as well as about the ethical implication which we will keep in consideration while executing the project

1.1. Exploratory Data Analysis / Data Cleaning

Before beginning the data analysis, we will understand the data first and more important we will need to understand the structure of cases. The first subheading in the case is introduction which consists the facts and point of appeal which suggest why the case come to the court. The second subheading is the evidence which consist testimonials of witness in the court. The third subheading is the appeal which consist the discussion related to this case at the court. In the appeal section we can able to find the case law which is used in the case. The final heading of the case is review of sentence in which outcome of the case is mentioned. We can use the case law from west-law database or lexis database which are popular databases containing huge amount of the case. The format of case is more same as we discussed earlier. We can download these cases from this database and can create a model using natural language processing. We can extract the appeal section from all cases which will help to generate the model faster, if we passed the whole document for processing then it can take time for processing, so to avoid this we can implement the extraction of appeal sections from all cases in the dataset which will be the first step of data cleaning. Now we can proceed with the basic data cleaning of the text data. In the first step, we can remove the punctuation because these punctuation marks does not make sense during the analysis of the words. In next step, we can convert all the text in the documents to lower case which will be helpful while doing future analysis. Further we can remove the stop words such as a, an, the, is, for and along with the stop word we can also remove the unwanted text such as symbols which does not make much sense. In next step we can implement the tokenization which is categorised as words tokenization and sentence tokenization. In word tokenization each word will separated and in sentence tokenization sentence separated by the full stop will be separated. Next step will be executing lemmatization and stemming in which words are converted into root words for example appealing is converted into an appeal. Stemming trim such words and remove the ‘ing’ and ‘ed’ at the end of the words to convert them into a root word whereas lemmatization does not work on the principle of removing key words such as ing or ed rather it looks on word vocabulary known as word net and it always gives meaningful words. These are the processes we will need to implement for data cleaning.

1.2. Dimensionality Reduction / Feature Selection

Now the process of modelling starts, so suppose there are three types of cases in the dataset which are criminal case, civil case and family cases. If we train the data based on these classes then model will able to predict the document belong to which class. For fitting the data in the machine learning model we will need to execute the feature extraction process. To begin with feature extraction we will need to understand two terms which is corpus and bag of words. Corpus is the combination of all text data present in all documents of training dataset. Bag of words will contain the unique words contain in the corpus but the order of sentence is not preserved in the bag of words. Further will create a vectors of our training documents which can be fed in to a machine learning model. Vector is produced with help of bag of words and check or count the frequency of each bag of words in a document. We can use count vectorizer package in the python to implement this step. Further we can use TF-IDF which stands for term frequency-inverse document frequency. TF-IDF is used to supress the effect of repeating words in the document. In our case legal can be the word we can find in all three types of cases such as criminal, civil and family. We will need to supress effect of word legal in the vector and we to increase the effect of criminal, civil and family, term frequency is used.

1.3. Feature Engineering / Feature Extraction

Bag or words and TF-IDF is similar to one hot encoding and to improve this technique further we can use word embedding. Word embedding is a matrix in which first column consists the features and the first row consist words and these matrix table will suggest the feature relation value of the words for example in our case words such as affidavit, attestation are the words which are only found in civil cases and not in criminal case. So such words will have high score for civil feature and low score for other feature like criminal. In this method we will have low dimension and dense matrix.

1.4. Choice of modelling techniques

While selecting the modelling technique, I was confused about selecting one technique among the recurrent neural network and convolutional neural network. After studying about both algorithms I have found that recurrent neural network cannot capture phrases without prefix context and it also often capture too much of last words in final vector. For using convolutional neural network we will need to implement the bigram model or trigram model because this will help network to learn neighbouring representation and then join them together level by level. Let us first understand the CNN implementation on the image data first. In an image data, image is represented by a matrix which is the shape of number of pixel for example 200 x 200 pixels and then a filter is applied on this matrix and what we get at the output of this operation as convolved feature. Similarly in text processing, we have created a word embedding matrix which is similar to pixel matrix of an image. But in text processing there will be a one dimensional convolution. We can apply filter or feature map of particular size, suppose if we take filter size of 3 so it will take the three vectors and apply convolution of feature matrix with word embedding vector. It will give a one vector at the output which will represent the feature map for the whole sentence. During these convolution process the first and last row will be neglected due to the kernel size. In order to include these rows we will need to implement the padding. We can pad zeros at first row and also at last row so that we will not lose the length of the vector. The size of the filter is important because it suggest how many number of neighbouring words you are considering for this filter. We can also apply multiple filter on the word embedding vector and number of vectors at the output will be equal to the number of filters applied. Usually, if we use more number of filters then there are more chance of extracting more data and model can perform better. Similar to convolutional neural network for images, in text classification too we can add max pooling layer which will store the maximum value of each vector generated by each filter after convolution. After max pooling, we will get the vector of length equal to the number of filter applied. Yoon Kim (2014) implemented the sentence classification using convolutional neural network. The goal of this project was to classify the sentiment of the sentence. Also the other application of this project was to classify whether a sentence is subjective or objective and to classify the question whether sentence ask about a person or location or number.

1.5. Hyperparameter Optimisation

There are many hyperparameters in convolutional network which can be adjusted during modelling. We can adjust the stride of filter which suggest filter to move how many steps at a time. We can do k-max pooling in which k vectors can be generated after the pooling layer. We can also add a dropout which is useful for regularizing the output. In dropout layer few neurons are deactivated, this method helps to avoid the overfitting of the model. Another way to regularise the model is to just constrain the norms of weights vectors to same scale. Yoon Kim (2014) used the following hyperparameters in their model. Researcher used the relu activation function and used three different size of the filter (h=3, 4, 5). Each filter had 100 feature maps. Researcher noted that after adding the dropout layer there was improvement in the accuracy by 3-4%. For regularization researcher used L2 constraint and along with this they used training batch size of 50. Also researcher used the pretrained word2vec of dimension equals to 300. Also we can use the batch normalisation in which output is scaled from the layer for that batch to all have zero mean and unit variance. Using batch normalization, models produced are less sensitive to parameter initialization and also make tuning of learning rates simpler.

1.6. Model Evaluation

After producing a convolutional neural network we will need to see how well the model performs. We want to see how good the outputs are and as well as we will need metric to compare our model with previous models. We will evaluate the model based on accuracy, precision, recall and f1 score. These performance metrics are derived from classification matrix which is divided into four parts. True positive correctly identifies the prediction of each class. True negative is correctly rejected the prediction for certain class. False positive means incorrectly identified predictions for certain class. False negative means incorrectly rejected data for certain class. Accuracy is calculated as the total number of correct predictions divided by the total number of training data. But accuracy can mislead us while selecting the better because if we consider our data is imbalanced then model can tend to predict more values correctly of class which have large data whereas the model correctly classifies only few cases for the classes which have small data. In this scenario the overall accuracy of the model will increase but the accuracy of model for predicting the class which have less data will be too low. So to measure the performance of imbalanced data f1 score performance metrics is used. F1-score considers both cases like for a given case will classifier able to detect (recall) and for a class prediction from classifier, how likely is to be correct (precision). F1 score is the harmonic mean of the recall and precision. Harmonic mean is used to punish the extreme values more so this method is useful for imbalanced data.

1.7. Scalability Issues

The word2vec model was implemented by programming language like C which was trained on single CPU. Also the initial thinking about the text data that it will not require the GPU for developing the model but it is incorrect. We will be dealing with deep learning method in our project so to make the process fast and efficient we will need the GPU utilisation. In our scenario we are assuming the corpus of legal words will be large so to train the vectors based on this corpus we will need GPU to make this process faster.

1.8. Ethical implications

We will be using the west-law or lexis database to download the past cases. These two databases are public database. In our case, there cannot be any issue on privacy like in case data, there are no records of personal information of judge or appellant. As we will be picking the random cases which will neglect the bias in the data. We will need to use the updated version of each cases which we will be using on training and testing.

2. Question 2 – Paper Review

The second part of the assignment is related to a paper review of the article “Wheat varieties identification based on a deep learning approach ” produced by Karim Laabassi et al. (2021).

2.1. Structure and Title

The structure of the paper is quite standard, they first covered the abstract which suggested why research was important in their area and what methodology they will covering and as well as what results they achieved in their research. Further in the introduction they introduced about the quality inspection methods of wheat. They described their methodology in material and method. They described about the result briefly in the results and discussion section. They concluded their research with by summarizing their findings. After reviewing the whole research paper, I can suggest the tile of paper should have been Wheat varieties identification based on transfer learning approach.

2.2. Abstract

In the abstract they introduced about wheat recognition and why their method is novel and useful for wheat recognition. Further they described about the methodology they are going to use which was transfer learning. Their main purpose of the research was to calcify four varieties of the wheat (Simeto, Vitron, ARZ, and HD). They concluded their abstract by describing the results of each model. Abstract covered all important aspects of the article.

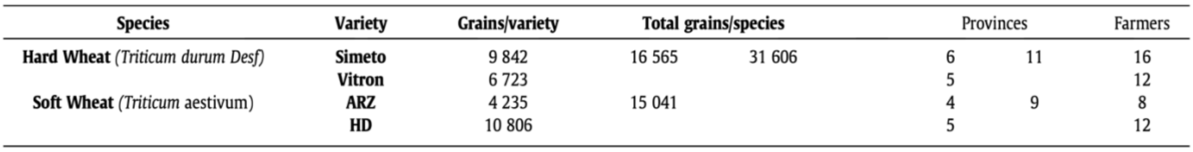
2.3. Introduction and Previous Research

Researchers initially discussed about the quality testing of wheat crop. They have stated their objective here. Further they discussed about the grain identity recognition (GIR) which is used for seed tester and through this research they know about the two level of classification which is species level classification (SLC) for physical purity test and varietal level classification (VLC) for varietal purity. Researchers discussed about the varietal level classification and analysed why it is difficult to implement the VLC due to the similarity between different species of the grains. Further they cited the research in which grain identification was achieved using computer vision and deep learning and researchers suggested that this method can help to perform VLC accurately. Here researchers stated their hypothesis. Further they studied few research papers in which image classification and machine learning is used for cereals and for automatic grain classification. Researchers discussed about the image classification approaches, in shallow classical approach a hand crafted feature extraction is done for the input data and it is then passed for the trainable classifier. In deep learning approach in which images of different categories are directly used by the deep learning classifier. They discussed the drawbacks of shallow classical approach by reviewing few studies in which neural network was used and they got to know that that neural network failed to classify the objects which had little difference in colour, textural features etc. Further they studied about the deep learning architecture by reviewing few researches in which classifier was produced from the scratch and understood the computation requirement for deep learning approach. Further they also studies about the convolutional neural network (CNN) and recognized CNN as best architecture for deep learning. They discussed about the few studies in which CNN was used in transfer learning to identify the insect species. They studied about the transfer learning in detail with the help of few studies and understood that large image dataset can be trained using transfer leaning to successfully classify the images or objects. They studied about the Fast R-CNN which was used for insect identification in stored grains. Researchers discussed about two studies whose topic was close to their research. In that research, varietal level classification of barley was achieved using convolutional neural network. They also used five transfer learning model in their research. In second study they cited, they understand the effect of using balanced dataset on the accuracy as well as researchers got the idea for data collection using flatbed scanners. Further researchers discussed the limitations or drawback in data acquisition step in previous researches. They discussed about the segmentation and with the help of segmentation how CNN can focus only on the object and not on the background of the image. They had provided the comprehensive information about the past researches. Researchers concluded the introduction by stating how their research is novel based on the studies they discussed also researchers discussed about novel approach of data collection in which they collected grains from different regions and performed two way sampling. They basically carried out the literature review in the introduction section itself due to the guideline of Journal of Saudi Society of Agricultural Sciences.

2.4. Methodology

Researchers divided their methodology section in two parts. In first part, they explained the process of data collections and in second part they explained about deep learning classification implementation. They had provided the sampling information in detail in their paper and also they created a process flow in which they showed the data collection steps which helps to understand the process much better. They classified the wheat in two categories which are hard wheat and soft wheat and again these each categories are further categorized in two categories as Simeto, Vitron, ARZ, and HD respectively. In sampling, they did not considered any fixed ratio of samples per category and sampling was done randomly. Their helped us to understand the sampling of grain better.

Table 1. Sampling of Grains



Species	**Variety**	**Grains/variety**	**Total grains/species**		Provinces		Farmers
Hard Wheat *(Triticum durum Desf)*	**Simeto**	9 842	16 565	31 606	6	11	16
	Vitron	6 723			5		12
Soft Wheat *(Triticum* aestivum)	**ARZ**	4 235	15 041		4	9	8
	HD	10 806			5		12

They have used transfer learning approach in which they added the global average pooling layer which averages the feature maps present in the last layers of the models like InceptionV3, MobileNet, Xception, ResNet50, and DensNet201. This step reduces the complexity of the model and also reduces the risk of overfitting. Also I found that they could have used data augmentation technique to make the model better for predictions. They have added one input layer, one global averaging pooling layers, and a dense layer with SoftMax function. They could have added a dropout layer after the pooling layer. Researchers explained the methodology in correct manner by explaining the data collection step first and then they explained about the implementation of models and due to the use of appropriate figures and tables it was easy to understand the methodology

2.5. Results

Researchers started the result section by comparing the validation accuracy of five models and arranged these model in descending order with respect to validation accuracy. As they produced the models using transfer learning method, so sometimes it is obvious that training score is high so they made more emphasis on discussing the test scores of these five model. Again, they used table format to show their results which helps better to understand the results. They compared the validation and testing score and noted their observation of decrease in accuracy of model during testing by 3-4% which is quite low. They have also categorized the model depending upon the classification table and noted under the results which model is predicting respective categories of wheat at high accuracy. By using the classification tables, it was easy to interpret the result of the study. Further researchers discussed and compared their results with previous studies. So they summarized the results and methodology in which they proposed varietal level classification which was unique compared to previous studies and they used only dorsoventral side of the grain during this study. They compared their results with Kozowski et al. (2019) in which they used barley's for classification and noted their observations that they achieved high accuracy than this study. Further researchers compared another study in which fine tuning was used and noted that even not using the fine tuning method in this study how their results are much better than study in which fine tuning was used. Due to use of large dataset and each varieties having finite number of images so they did not required to use fine tuning.

2.6. Conclusion/Discussion

Researchers concluded their research by suggesting how their model can be implemented in the real world scenario. They suggested the Mobile-Net model can be easily used in the intelligent embedded devices. During the review of paper, I have observed the training and validation accuracy and loss graph and the peaks in the graph was noticeable which suggest that final batch in a epoch could have small data or Batch Norm with small batch size and large epsilon ϵ (hyperparameter) but they provided the reason in the conclusion that because of not using regularization techniques few peaks were generated in the loss graph. Researchers also concluded that their study contributed in phenotyping sector and in future work it can be used for phenotyping and as well as genetics field.

Reference

M. Kozowski, P. Gœrecki and P. M. Szczypiski. "*Varietal classification of barley by convolutional neural networks*". Biosyst. Eng. 2019, pp. 155–165, doi: 10.1016/j.biosystemseng.2019.06.012.

K. Laabassi, M. A. Belarbi, and S. Mahmoudi , "*Wheat varieties identification based on a deep learning approach*", Journal of the Saudi Society of Agricultural Sciences, 2021, pp. 1-9, doi: 10.1016/j.jssas.2021.02.008

Kim, Yoon, “*Convolutional Neural Networks for Sentence Classification*”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, doi: 10.3115/v1/D14-1181.