

***Application to Register for a Postgraduate Research Programme***

(To be completed in consultation with the Director of Studies)

*In completing this form, the applicant should refer to the relevant sections of the Part 9 of the Manual of General Regulations, the Code of Practice for Postgraduate Research Degrees and, where necessary, the Code of Practice for Research Ethics and Governance.*

***This form must be signed and dated in advance of submission to School Research Degrees Sub-Committee (SRDSC).***

***What Now?*** *This form should be completed by the applicant in consultation with the Supervisory Team. Following formal approval at Research Degrees Subcommittee, written confirmation will be sent to the applicant from the relevant officer within the academic school.*

**1. Student’s Details**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Full name** | **Chung Hoi CHIU** | | | | |
| **UEL student number** | **U1720146** | | | | |
| **Programme for which you are currently Enrolled** *(Please Tick)* | **MPhil** | | |  | |
| **MPhil by Published Work** | | |  | |
| **PhD via MPhil** | | |  | |
| **PhD Direct** | | |  | |
| **PhD by Published Work** | | |  | |
| **Prof Doc** | | | **P** | |
| **Title of Professional Doctorate Programme (if applicable)** | **Data Science** | | | | |
| **Date of first enrolment for current programme** | **1 October 2019** | | | | |
| **Has the period of enrolment involved a suspension of studies?**  (Delete as Appropriate) | **~~Yes~~** | | **No** | | |
| **If the period of enrolment has involved a suspension of studies, please indicate the duration in whole semesters.** |  | | | | |
| **Current Mode of Study**  (Delete as Appropriate) | **Full Time** | | **~~Part Time~~** | | |
| **Number of Modules Currently Registered for** *(For Professional Doctorates only)* | **One** | | | | |
| **Requested date for start of registration** *(Registration must occur within 6/12 months of the date of enrolment for FT/PT students respectively and may be backdated to that date.)* | **April 2021** | | | | |
| **Declaration of previous registration.** (*If a previous period of registration exists elsewhere this may be transferred to form part of the registration period at UEL. Evidence confirming the dates and location of this registration must be provided.)* |  | | | | |
| **School** | **School of Architecture, Computing and Engineering** | | | | |
| **1) Name of Collaborating**  **establishments (if any)**  **2) Please list your**  **undergraduate and/or**  **postgraduate**  **qualifications** | **2) BSc (Hons, B’ham), MBA(HKU), LLM(London), MA(KUL), MSc (UEL)** | | | | |
| **Relationship between work to be undertaken in the collaborating establishment and that to be undertaken at the sponsoring establishment or elsewhere (if relevant)** | | | | | |
|  | | | | | |
| **Details of facilities available for the investigation, including funding and location** | | | | | |
|  | | | | | |
| **If the student will be domiciled outside the UK while the research is being conducted, please provide details of proposed arrangements for supervision, specifically the nature and frequency of contact*.* Please note, particular attention should be paid to the demands placed on candidates while not in attendance at UEL, and guarantee that appropriate supervisory support is maintained.** | | | | | |
|  | | | | | |
| **In the event that the student intends to undertake the majority of their research at a higher education provider or research institution abroad, have Split-Site arrangements been considered by School Research Degrees Sub-Committee (SRDSC)?** (Delete as Appropriate) | | **Yes** | | | **No** |

**2. The Programme of Research**

|  |
| --- |
| **Proposed Title of Thesis** |
| Approaching authorship attribution studies with artificial neural network methods |
| **Aim of the Investigation** |
| To invent novel and effective methods of artificial neural network and the corresponding data pre-procession methods for conducting authorship attribution studies. |
| **Details of proposed research in lay terms** |
| Traditional machine learning methods (ML) are already very well developed for resolving authorship attribution (AA) queries. For example, in 2013, *Sunday Time* commissioned the expert in forensic AA, Professor Patrick Juola, to study the true author of the novel *Cuckoo's Calling*. Within days, he identified J. K. Rowling is the true author, out of four ‘suspects’ (<https://languagelog.ldc.upenn.edu/nll/?p=5315> ). Last year (2020), in doing coursework DS7003 (<https://github.com/ericchchiu/u1720146_DS7003_courseworkCodeAndData> ), I correctly attribute 59 or 60 pieces out of the total of 60 4000-word texts to seven nineteenth century British novelists. However, traditional ML can use only for studying the frequency of features, and not their sequence and distribution. Therefore, academia and industry are keen on encouraging partitioners to research using the sequence and distribution of features to perform AA, including using the methods of artificial neural network (ANN, also known as deep learning methods). This research aims at finding novel and effective ANNs for conducting AA. |
| **Proposed plan of work, including its relationship to previous work, maximum 4,000 words.** *Please include in the discussion a description of the research methodologies and explain why these methodologies are the most appropriate for the task. Include a list of references for all works cited. Gantt charts should be included, where appropriate, to reflect research planning.* | |
| (Please see the attachment) | |

|  |
| --- |
| **Summary of the elements of the investigation that are novel, original or creative and that may constitute production of original knowledge or an original interpretation of existing knowledge.** |
| The research will lead to:  1. the production of novel and effective artificial neural network models (also known as deep learning models) for conducting authorship attribution studies.  2. the improvement of certain existing artificial neural network models which are used to conduct authorship attribution studies.  3. understanding more about the extent of the usefulness of linguistic information extracted from text data to the application of artificial neural networks to authorship attribution studies. |

**Risk Assessment and Overseas Travel**

***A Research Risk Assessment is required for any proposed laboratory experiments and/or fieldwork to be conducted off-site. Where the fieldwork involves overseas travel, separate approval from Vice-Chancellor’s group (VCG) must be obtained in advance of this activity.***

|  |  |  |
| --- | --- | --- |
| **Does this investigation require laboratory experiments and/or fieldwork?**  (Delete as Appropriate) | **~~Yes~~** | **No** |
| **If laboratory experiments and/or fieldwork are involved, has the Research Risk Assessment been carried out?**  (Delete as Appropriate) | **~~Yes~~** | **No** |
| **If the proposed research involves overseas fieldwork, has a request for approval of this activity, via form RFA(2), been submitted to Vice-Chancellor’s Group (VCG)?**  (Delete as Appropriate) | **~~Yes~~** | **No** |

***NOTE: Where the proposed research involves overseas travel for the fieldwork, the registration arrangements may be approved subject to the understanding that such activity may only be carried out with advance approval from Vice-Chancellor’s group (VCG) via form RFA(2)*** [***http://www.uel.ac.uk/qa/research/fieldwork.htm***](http://www.uel.ac.uk/qa/research/fieldwork.htm)

***The completion of a risk assessment for the research is the primary responsibility of the applicant and the Director of Studies. School Research Degrees Sub-Committee (SRDSC) should receive confirmation that the risk assessment is complete. Where is it apparent that the risk assessment has not been carried out, this should be reflected in the accompanying minute from SRDSC, clarifying required action to ensure its completion.***

**Research Ethics and Integrity**

|  |  |  |
| --- | --- | --- |
| **Does the proposed research design include parameters requiring ethical approval?**  (Delete as Appropriate) | **~~Yes~~** | **No** |
| **If ethical approval is required, has the requisite application been submitted to University Research Ethics Committee (UREC)?**  (Delete as Appropriate) | **Yes** | **No** |
| **Has the applicant sought ethical approval from an external REC in advance of submitting an application to UREC?**  (Delete as Appropriate) | **~~Yes~~** | **No** |
| **I have attached my certificate of completion of the research integrity module** | **Yes** | |

***Note: Where the research involves human participation (i.e. human participants, human material or human data) prior approval from a research ethics committee is mandatory. Applicants should be aware that conducting any form of research involving the above parameters without appropriate ethical approval may result in disciplinary action and/or the cancellation of the research.***

***Note: The Research Integrity Module is compulsory for all Postgraduate Research Students and proof of completion must be submitted with your completed registration document.***

**Intellectual Property Rights**

|  |  |  |
| --- | --- | --- |
| **Do the proposed arrangements include a transfer on to a UEL postgraduate research degree following registration with another Higher Education Provider?** (Delete as Appropriate) | **~~Yes~~** | **No** |
| **Will the programme of research lead to output(s) which will have commercial form(s) and/or intellectual property of potential value?**  (Delete as Appropriate) | **~~Yes~~** | **No** |
| **If output is of potential value, has the Office of Intellectual Property Development been consulted?**  (Delete as Appropriate) | **~~Yes~~** | **No** |

**3. Proposed Supervisory Team**

|  |  |
| --- | --- |
| **Nomination of Director of Studies (DoS)** | |
| **Name and title** |  |
| **School** |  |
| **Current position, Department and Institution** |  |
| **Previous posts held** |  |
| **Email address** |  |
| **Qualifications** |  |
| **Current research or professional practice** |  |

|  |  |  |
| --- | --- | --- |
| **Is the proposed supervisor currently registered as a student on a research degree programme at any HEI?**  (Delete as Appropriate) | **Yes** | **No** |

|  |  |  |
| --- | --- | --- |
| **Number of HEI research degree students currently supervised** | **MPhil** |  |
| **Professional Doctorate** |  |
| **PhD** |  |
| **Number of HEI research degree students successfully supervised to completion** | **MPhil** |  |
| **Professional Doctorate** |  |
| **PhD** |  |

|  |
| --- |
| **Please list up to six publications which are of most relevance to this proposal. If there are no publications, please indicate the relevant professional Experience.** |
|  |

|  |  |
| --- | --- |
| **Nomination of Second Supervisor** | |
| **Name and title** |  |
| **School** |  |
| **Current position, Department and Institution** |  |
| **Previous posts held** |  |
| **Email address** |  |
| **Qualifications** |  |
| **Current research or professional practice** |  |

|  |  |  |
| --- | --- | --- |
| **Is the proposed supervisor currently registered as a student on a research degree programme at any HEI?**  (Delete as Appropriate) | **Yes** | **No** |

|  |  |  |
| --- | --- | --- |
| **Number of HEI research degree students currently supervised** | **MPhil** |  |
| **Professional Doctorate** |  |
| **PhD** |  |
| **Number of HEI research degree students successfully supervised to completion** | **MPhil** |  |
| **Professional Doctorate** |  |
| **PhD** |  |

|  |
| --- |
| **Please list up to six publications which are of most relevance to this proposal. If there are no publications, please indicate the relevant professional Experience.** |
|  |

|  |  |
| --- | --- |
| **Nomination of Third Supervisor (if applicable)** | |
| **Name and title** |  |
| **School** |  |
| **Current position, Department and Institution** |  |
| **Previous posts held** |  |
| **Email address** |  |
| **Qualifications** |  |
| **Current research or professional practice** |  |

|  |  |  |
| --- | --- | --- |
| **Is the proposed supervisor currently registered as a student on a research degree programme at any HEI?**  (Delete as Appropriate) | **Yes** | **No** |

|  |  |  |
| --- | --- | --- |
| **Number of HEI research degree students currently supervised** | **MPhil** |  |
| **Professional Doctorate** |  |
| **PhD** |  |
| **Number of HEI research degree students successfully supervised to completion** | **MPhil** |  |
| **Professional Doctorate** |  |
| **PhD** |  |

|  |
| --- |
| **Please list up to six publications which are of most relevance to this proposal. If there are no publications, please indicate the relevant professional Experience.** |
|  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Overall supervisory experience and activity of the Proposed Supervisory Team** | | | |
| **Has every member nominated to act in a supervisory role completed the requisite UEL Supervisor Training Programme?**  (Delete as Appropriate) | | **Yes** | **No** |
| **Collective number of research degree students successfully supervised to completion** | **MPhil** |  | |
| **Professional Doctorate** |  | |
| **PhD** |  | |
| **If the combined experience and activity of the proposed supervisory team does not meet the requirements stipulated in UEL’s Research Degree Regulations, please provide a short statement justifying why consent is sought and why this particular supervisory team is most suitable for the programme of research.** | | | |
|  | | | |

**4. Nomination of Advisor**

|  |  |
| --- | --- |
| **Name and title** |  |
| **Current position, department and institution** |  |
| **Full Postal address** |  |
| **Email address** |  |
| **Previous posts held** |  |
| **Qualifications** |  |

**5. Student’s Declaration**

*please note that electronic signatures are NOT acceptable.*

|  |  |
| --- | --- |
| **I Confirm:**   1. **that I wish to apply to be registered as a student for the postgraduate research award as indicated above;** 2. **that the details given in this form are, to my knowledge, correct;** 3. **that any written component of the programme will be submitted in English;** 4. **that I agree to undertake oral examination in English;**   ***(Requirements 3. and 4. may be waived where separate arrangements are approved by Research Degrees Subcommittee)***   1. **That I have read and agree to the proposed supervision arrangements.** 2. **That I will obtain prior approval for any overseas fieldwork, in advance of the commencement of such activity, from Vice-Chancellor’s Group (VCG).** | |
| **Student** | **Signed:** |
| **Date:** |

**6. Supervisory Team’s declaration**

*please note that electronic signatures are NOT acceptable.*

|  |  |  |
| --- | --- | --- |
| **We confirm:**   1. **that we support this application and confirm that the Student has demonstrated the potential to complete the proposed programme of work;** 2. **that no member of the supervisory team is currently registered on a research degree programme at any HEI;** 3. **that, if applicable, we agree to the request for the backdating of registration;** 4. **That we recommend that the applicant be registered for a research degree Programme at UEL.** 5. **That we have briefed the applicant on all requirements relating to research ethics and research governance, relevant to the proposed research.** | | |
| **Director of Studies** | **Signed:** | |
| **Printed:** | **Date:** |
| **Second Supervisor** | **Signed:** | |
| **Printed:** | **Date:** |
| **Third Supervisor  (if applicable)** | **Signed:** | |
| **Printed:** | **Date:** |

**7. Dean of School’s declaration**

*please note that electronic signatures are NOT acceptable.*

|  |  |  |
| --- | --- | --- |
| **I confirm that our University and School facilities and resources detailed in this form, including academic staff involved in the student’s supervision, will be available for the duration of the programme of research.** | | |
| **Dean of School** *(or nominee)* | **Signed:** | |
| **Printed:** | **Date:** |

**Attachment**

***PROPOSED PLAN OF WORK, INCLUDING ITS RELATIONSHIP TO PREVIOUS WORK, MAXIMUM 4,000 WORDS. Please include in the discussion a description of the research methodologies and explain why these methodologies are the most appropriate for the task. Include a list of references for all works cited. Gantt charts should be included, where appropriate, to reflect research planning.***

**1. Introduction**

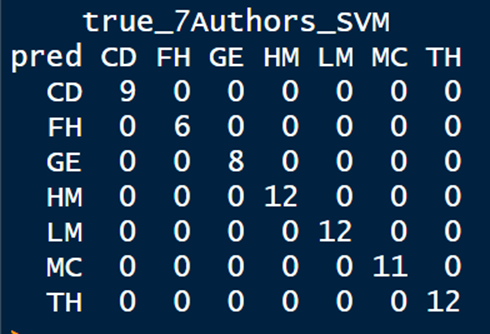
The two paragraphs below are quoted from my CN8002 coursework (Chiu, 2019), which shows how the discipline of computational authorship attribution (AA) began:

Computational authorship attribution is a discipline of study that uses computers to determine who wrote a piece of text the author of which is disputed or unknown. This discipline began in 1964. In that year, with the permission of the United States Government, the two pioneers, Federick Mosteller and David Wallace, reported in their book *Inference and Disputed Authorship: The Federalist* (Mosteller and Wallace, 1964, p. x) that they and their team studied texts of about 210,000 words that were typed on punch cards with computers located at the M.I.T. Computer Center, Cambridge, Massachusetts, and at the Center for Advanced Study in the Behavioral Sciences, Stanford, California respectively and found that the actual author of all of the 12 disputed Federalist Papers is James Madison, the fourth president of the United States (the Federalist Papers consist of 85 papers. The disputed ones are Nos. 49-58 and 62-63). Mosteller and Wallace used Bayes's Rules to make statistical inferences on log-negative binomial distribution probabilities (or on the simpler but less accurate log-poisson distribution probabilities) of the 30 carefully selected marker words (examples of the words are while, whilst, on and upon). The conclusion is convincing. Since 1964, dozens of different methods have been invented to study the 12 disputed papers, and most of them obtained the same or slightly deviated results. *Non-Stop*, one of the songs contained in the musical play *Hamilton* which has been performed in the West End (Victoria Palace Theatre) since 2017, has the following line: “John Jay got sick after writing five. Madison wrote twenty-nine. Hamilton wrote the other fifty-one.” The numbers contained in this line are based on Mosteller and Wallace’s research results.

Since the publication of Mosteller and Wallace's ground-breaking work in 1964, hundreds of works on this discipline have been published. According to Joseph Ruman, until 1998, more than 300 were published (Rudman, 1998); between 1998 and 2010, more than 600 were published comprised of journal articles, books, book chapters, dissertations, newspaper articles, on-line self-published papers and encyclopaedia entries (Rudman, 2010).

Then, in July of 2013, *Sunday Time* commissioned the forensic authorship attribution expert, Professor Patrick Juola, to study the true author of the novel *Cuckoo’s Calling*. Within days, Professor Juola identified J. K. Rowling as the true author, out of four ‘suspects’ (Juola, 2013a and 2013b).

Last year (2020), in doing coursework DS7003 (Chiu, 2020a), I correctly attributed 59 or 60 pieces, out of the total of 60 pieces of 4000- word texts to seven nineteenth century British novelists:



(Copied from Chiu, 2020a)

Both Professor Juola and me used traditional machine learning methods (ML) to conduct AA. Their achievements demonstrate that traditional ML are already very well developed for resolving AA queries. However, traditional ML can use only for studying the frequency of the features and not their sequence and distribution. For example, traditional ML would treat the sentence:

The quick brown fox jumps over the lazy dog.

The same as the following sentence:

The quick dog jumps over the lazy brown fox.

Similarly, traditional ML would treat ‘Mary likes John’ the same as ‘John likes Mary.’

Therefore, academia and industry are keen on encouraging partitioners to research using the sequence and distribution of features to perform AA, including using the methods of artificial neural network (ANN, also known as deep learning methods). For example, in the PAN of CLEF (2020) competition, a worldwide authorship attribution competition/ shared tasks, the organiser relaxed the computing power limitation and provided an additional large dataset, to encourage participants to use ‘data-hungry deep learning algorithms’ (PAN: Authorship verification 2020, 2020).

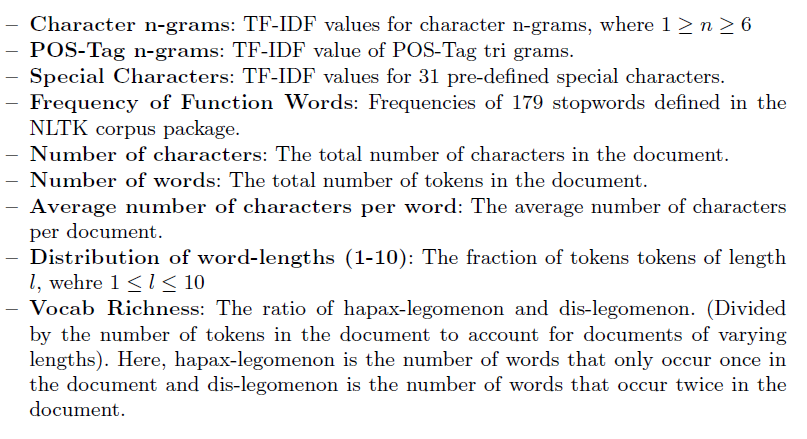
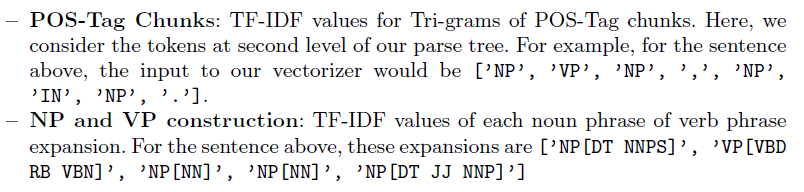
My research will focus on finding novel and effective ANNs to conduct AA. The remainder of this plan will be arranged as follows: in Section 2, I will describe the current situation of research on using ANN to conduct AA; in Section 3, I will describe the methodologies that I will use to conduct my research; and, in Section 4, I will provide a Gantt chart.

**2. Existing work on applying ANN to AA**

Currently, roughly speaking, practitioners researching the application of ANN to AA can be divided into three cohorts. They are described below:

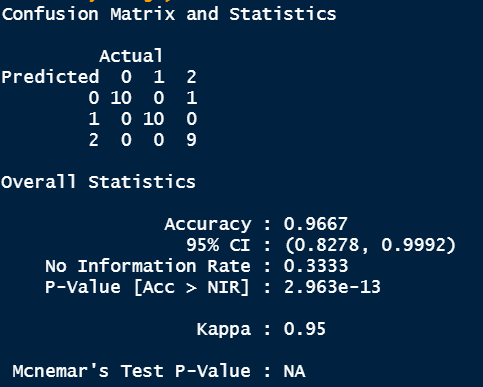
**2.1 Those who treat** **an ANN just like another ordinary ML classifier**

This cohort is represented by the runner-up of the PAN of CLEF (2020) competition (mentioned in Section 1), the weerasinghe20 team (Weerasinghe and Greenstadt, 2020). The team first extracted 11 features (shown on the below picture) from the large dataset and from the small dataset. The team then used a simple ANN to study the large dataset and used logistic regression, a traditional ML, to study the small dataset. The team thereby stuck to the traditional practice of paying attention only to the frequency of the features and ignoring their sequence and distribution. They did not have to use ANN. If they preferred, they could have used logistic regression, or any other traditional ML, to study the large dataset.

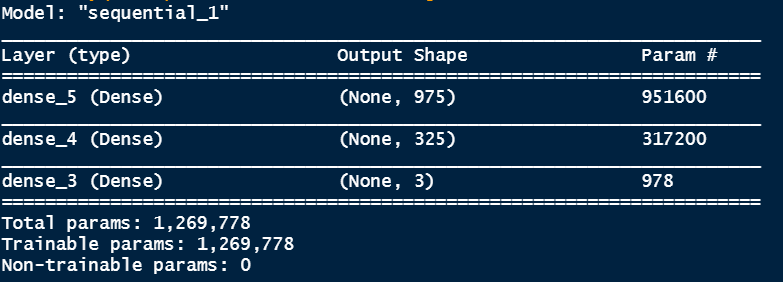
 

(Copied from Weerasinghe and Greenstadt, 2020)

When doing my DS7004 coursework, I also tried to use ANN in this way. The results, as shown below, were satisfactory. The ANN correctly attributed 29 out of the total of 30 pieces of texts to their authors:



Deployment of the ANN:



(The code for producing the above two pictures can be downloaded from: <https://github.com/ericchchiu/u1720146_DS7004_additionWorkChar4GramDNN> )

A careful study of the above 11 features extracted by the weerasinghe20 team reveals that they are unrelated or only remotely related to the topics the texts concern. This is a common practice in the field of AA (Goldberg, 2017, p. 78); paying substantial attention to features or words (i.e., the so-called content words) which are related to the topics that the texts concern, may end results in, for example, always thinking that two job application letters are written by the same author, and a job application letter and a tweet are written by two different authors. I advocate this line of thought and consider that, in most circumstances, content words are ‘noise’ in AA. But the second cohort of practitioners which I will deliberate in Section 2.2 think otherwise.

**2.2 Those who use ANN to extract linguistic information from data and then use the information** **to improve the model that was originally developed by using the whole of the training data, including content words**

From 2017 to October 2020, four arXiv papers were published concerning applying ANN to AA (Sari et al., 2017, Zhang et al., 2018, Jafariakinabad and Hua, 2019, and Jafariakinabad and Hua, 2020). The first paper claimed that it produced state-of-the-art results, because the results that it obtained were better than the results obtained by other practitioners who used traditional ML. The second and the third papers each claimed that each produced state-of-the-art results, because the results were better than the previous paper. The last paper was written by the same authors who wrote the third paper, and the last paper stated that the results it produced were better than the previous paper.

All four papers used ANN and treated topics and therefore content words of the data are useful for discerning who wrote the texts. The latter three papers also heavily relied on linguistic information extracted from data.

These four papers that consider topics of texts and therefore content words are useful for AA, just because both the training data and the test data they used were obtained from the same source and they never applied the trained model to texts obtained from a new source. For example, in Jafariakinabad and Hua, 2020, the following Python lines are used to pop out randomly 500 sentences from the training data to form the test data:

val\_indices = 500

…

…

val\_indices = np.random.choice(len(train\_lex), size=self.val\_size, replace=False)

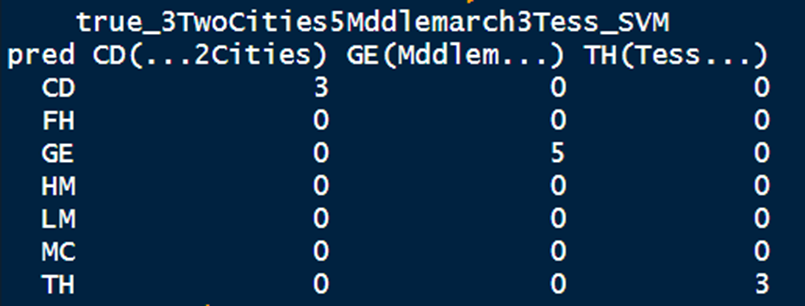
for index in reversed(sorted(val\_indices)):

val\_lex.append(train\_lex.pop(index))

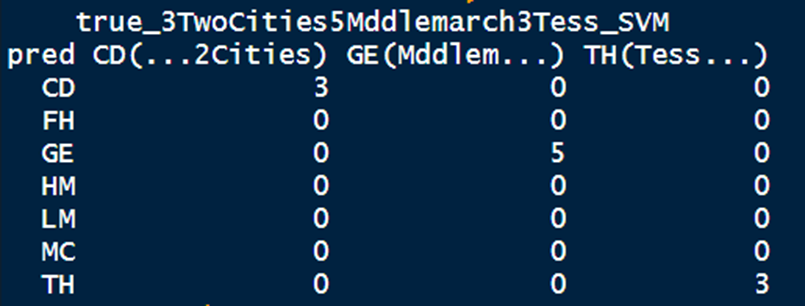
val\_synt.append(train\_synt.pop(index))

The paper does not consider what the results would be if the trained model is used to study texts that are not obtained from the same source from which the training data and the test data were obtained.

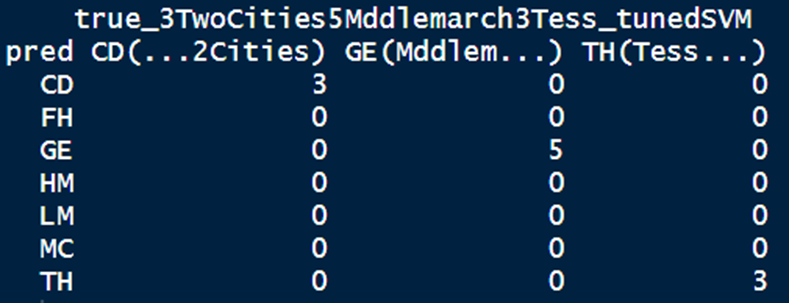
This is not a good practice. Therefore, in doing DS7003 coursework (Chiu, 2020a), after finding that my trained models can correctly attribute 59 or 60 4000-word texts to seven nineteenth century British novelists, I visited another text source (<https://www.gutenberg.org/>) to obtain a novel for each of Charles Dickens, George Eliot and Thomas Hardy: *A Tale of Two Cities*, *Middlemarch* and *Tess of the d'Urbervilles*. Texts of these three novels are not included in the training data or the test data. I then used texts extracted from these three novels to test whether the trained models still can correctly make determinations. The results are shown below:



Method: KNN



Method: SVM(turned automatically)



Method: SVM (turned manually)

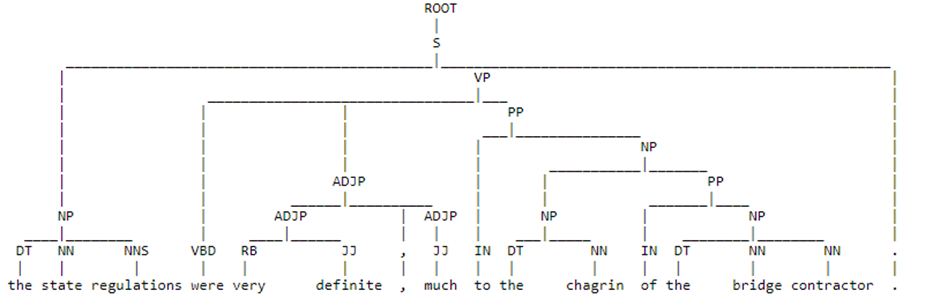
(The above three pictures were copied from Chiu, 2020a)

The models can still make correct determinations.

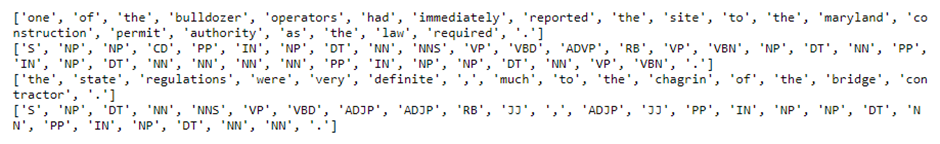
As the results obtained from the first of the four papers are not sufficiently accurate and the approaches of the remaining three papers are similar, hereinbelow I describe in relative detail only the methods used by the last paper (Jafariakinabad and Hua, 2020), which was published in October 2020:

The paper first split a text dataset into sentences (14 million+) and then used Stanford CoreNLP jar API and Python’s nltk package to form, for each sentence, a linearised parse tree. An example of how to form a linearised parse tree is shown below:

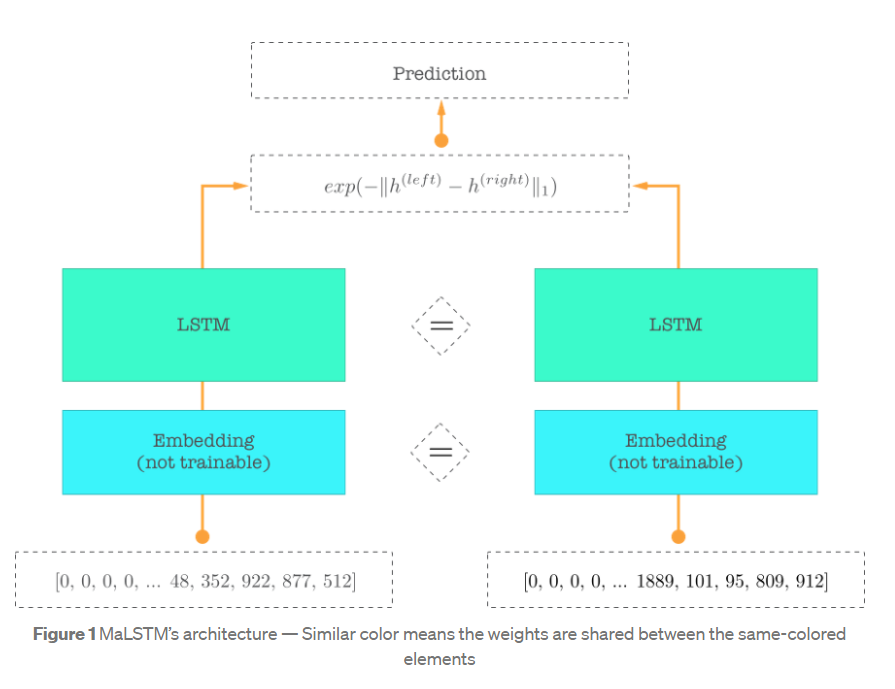
The sentence ‘the state regulations were very definite, much to the chagrin of the bridge contractor.’ was first converted to a parse tree, as shown below:



The parse tree was then flattened into a linearised parse tree, as shown below (at line 4):



The sentences and the linearised parse trees are then fed into an ANN invented by the authors of the paper and named the Lexicosynt network. Since the Lexicosynt network is similar to the Siamese network, I obtained a diagram of the Siamese network from the web for facilitating this explanation:



(Copied from Cohen, 2017)

The Lexicosynt network has the following three salient differences from the above Siamese network:

i. In the Lexicosynt network, the input is not pairs of sentences, but pairs of a sentence and a linearised parse tree. If the linearised parse tree is generated from the sentence of the pair, the label given to the pair is 1; if the linearised parse tree is randomly selected from the 14 million+ linearised parse trees and is not the linearised parse tree generated from the sentence of the same pair, the label given to the pair is 0. The pairs and the labels are produced programmatically, which is why the title of the paper is ‘A Self-supervised…’.

ii. In the Lexicosynt network, the two embedding layers are not pretrained and have different dimensions. According to the paper, the dimension of the embedding layer for sentences it set was 300, and the dimension of the embedding layer for linearised parse trees it set was 100. Both embedding layers are trainable.

Training this model is not to find whether the linearised parse tree of a pair of input was generated from the sentence of the same pair, but to obtain the trained 300-dimension word embedding layer. This is a novel way to produce a trained embedding layer. For a brief description of how to produce an ordinary embedding layer, please see my DS7004 coursework (Chiu, 2020b).

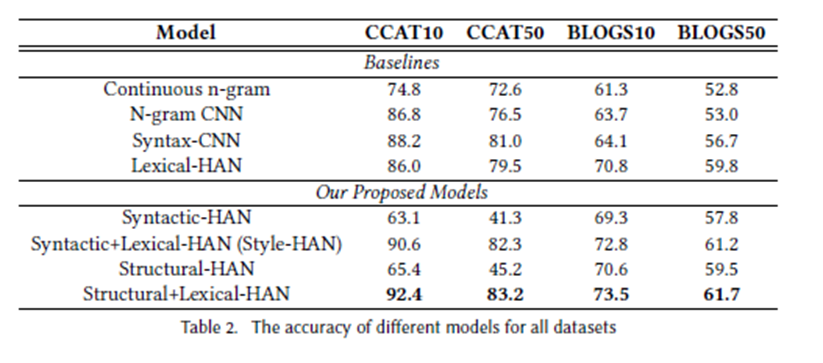
In an ordinary pretrained embedding layer, such as Facebook’s fasttext, Stanford’s gloVes or Gensim’s word2vec, location of a word’s vector representation is in accordance with the word’s lexicon meaning. Therefore, for example, if we obtain vectors of the words ‘king’, ‘queen’, ‘man’ and ‘woman’ from a pretrained embedding layer and perform calculation on them, we may find:

king – man + woman =~ queen

On the other hand, in the embedding layer obtained from training a Lexicosynt network, words are located according to their syntactic properties.

The vectors of the embedding layer obtained from training a Lexicosynt network (300-dimension, for example) are then concatenated to the vectors of an ordinary pretrained embedding layer (300-dimension, for example) to form, for each word, a 600-dimension vector. According to the paper’s studies, when conducting various types of NLP studies (Example of such studies may include those following the ‘four-step formula: embed, encode, attend, predict’ (Honnibal, 2016 and PyData, 2017)), including AA, using this combined embedding layer will, in most circumstances, provide better results than using just an ordinary pretrained embedding.

The paper provided a comparison table of the results obtained from using different methods:

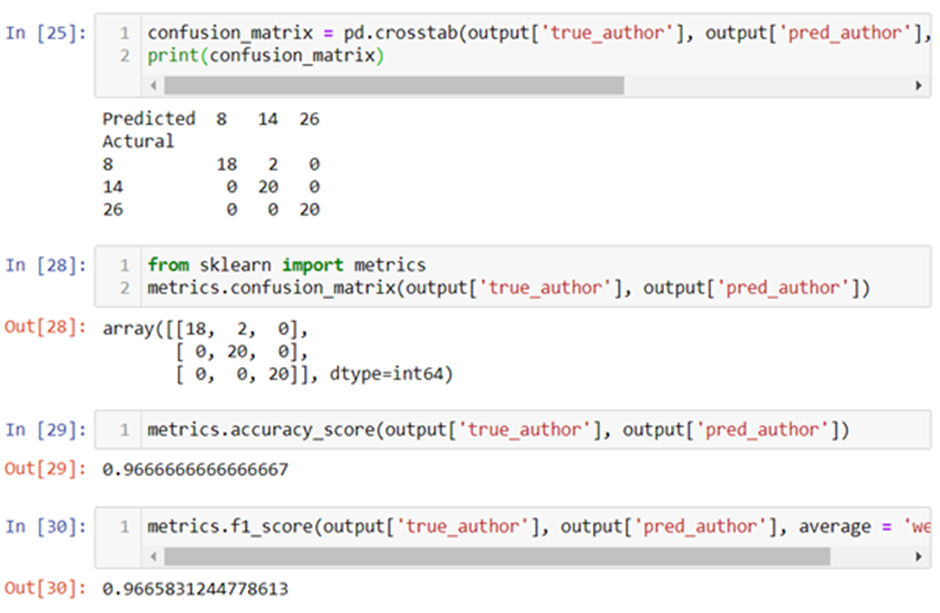


(Copied from Jafariakinabad and Hua, 2020)

The last line was obtained from using the method described above (i.e., using the combined 600-dimension embedding layer); the penultimate line was obtained by using just the 300-dimension embedding layer obtained from training a Lexicosynt network; the fourth line was obtained by using just fasttext, the popular pertained embedding layer developed by Facebook; and the third line was obtained by using the method provided by the second paper.

By adding the embedding layer trained by a Lexicosynt network to the ordinary pretrained embedding layer fasttext, the results did not improve very substantially (please compare the fourth line and the last line).

In doing coursework DS7004 (Chiu, 2020b), I also have tried to use a pretrained embedding layer (Gensim’s word2vec) to perform AA on texts of Charles Dickens, George Eliot and Jane Austen. I obtained an accuracy rate of about 85%. Then, for improving the results, instead of forming a special embedding layer by using a complicated ANN and a corpus which consists of 14 million+ sentences as described above, I just suppressed the influence of the content words and obtained the following results:

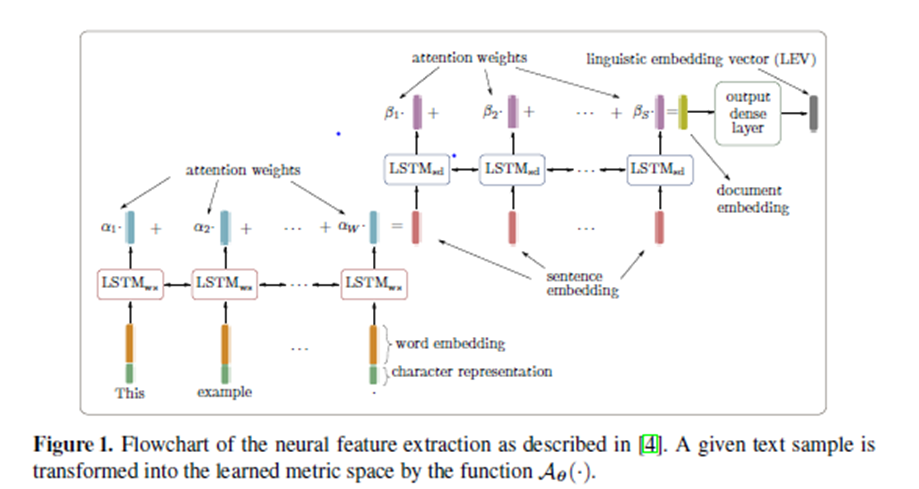


(Copied from Chiu, 2020b)

I increased the accuracy rate to 96%.

**2.3 Those who aim to designing a complicated ANN which can perform end-to-end process, from receiving raw data to producing results**

This cohort is represented by the winner of the PAN of CLEF (2020) competition (mentioned in Section 1), the boenninghoff20 team. This team designed a very complicated ANN to participate in the competition. The diagram of the ANN is shown below:



(Copied from Boenninghoff, et al., 2020)

The team called this ANN the ‘ADHOMINEM system’. ADHOMINEM is the abbreviation for ‘Attention-based Deep Hierarchical cOnvolutional siaMese bIdirectional recurrent nEural-network Model.’ The name of the model contains names of several cutting edge or trendy techniques in the field of ANN. They include attention mechanism, hierarchical neural network, convolutional neural network (CNN), Siamese network and bidirectional recurrent neural network (RNN).

In the team’s article summarising their methods (Boenninghoff et al., 2020), they expressly stated (bold added by authors of the article):

Automated (machine-learning-based) systems have traditionally relied on so-called **stylometric features**. Stylometric features tend to rely largely on linguistically motivated/inspired metrics. The disadvantage of stylometric features is that their reliability is typically diminished when applied to texts with large topical variations.

Deep learning systems, on the other hand, can be developed to automatically learn **neural features** in an end-to-end manner. While these features can be learned in such a way that they are largely insensitive to the topic, on the negative side, they are generally not linguistically interpretable.

The above quotation indicates that:

i. The team did not extract linguistic information, such as POS tags, character n-grams (used by the weerasinghe20 team, please see Section 2.1) and parse trees (used by Jafariakinabad and Hua, 2020, please see Section 2.2).

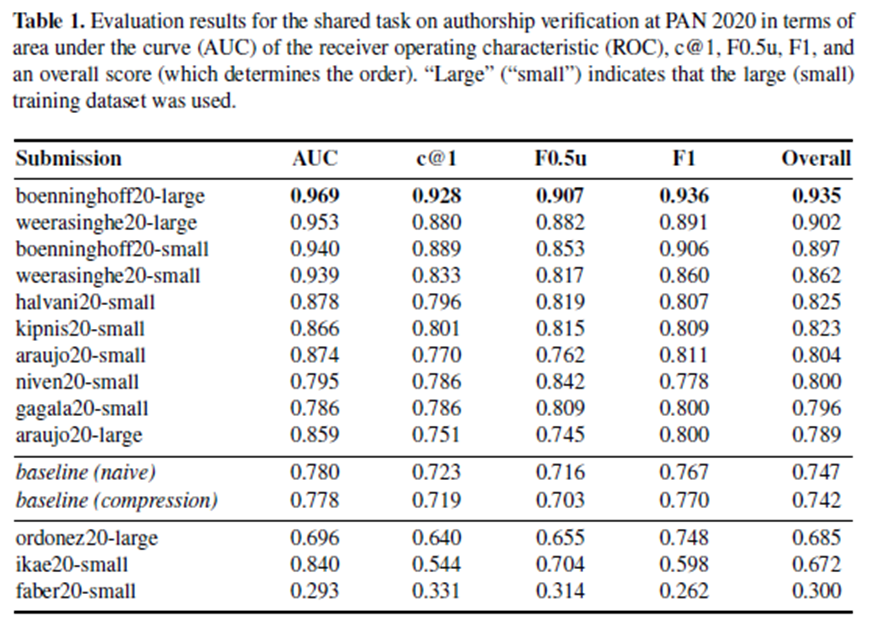
ii. The team focused on developing a complicated ANN that can learn ‘neural features’ from the raw data fed into it. My understanding is that ‘neural features’ mean features understood by and useful to the network, not humans.

iii. The team viewed words related to topics of texts (content words) to be a ‘noise’ for conducting AA.

The team still has not published the code used. But, the team said, ‘The source code will be publicly available to interested readers after the peer review notification, including the set of hyper-parameters.’ (Boenninghoff et al., 2020) I am waiting for the code before diving into investigating the team’s ADHOMINEM system.

The ADHOMINEM system consumes a large amount of computing power. It took six hours for the organiser of the competition to run the model submitted on unseen data. I found at the end of the article summarising the team’s methods (Boenninghoff, et al., 2020) the following acknowledgement: ‘This work was in significant parts performed on a HPC cluster at Bucknell University through the support of the National Science Foundation’ (HPC is the acronym for ‘high computing performance’). In contrast, the ANN model submitted by the runner-up of the competition, the weerasinghe20 team, required the organiser to spend only 2 hours and 19 minutes to run (Weerasinghe and Greenstadt, 2020).

The result table of the PAN of CELF competition (2020) is shown below:



(Copied from Bevendroff et al., 2020)

The table above shows that the runner-up of this competition, the weerasinghe20 team, did not lag the winner, the boenninghoff20 team, by much. The runner-up also used a much simpler ANN, which consumed much less computing power. A careful study of the features used by the runner-up (Section 2.1) will show that the runner-up can make certain easy improvements. For example, the team can drop those features that are widely agreed as not quite useful, such as the average total number of characters per document; or give more weight to those features that have proved very useful, such as character n-grams or frequency of function words.

**3. The methodologies I will use to conduct my research**

**3.1 A summary of the existing situation of applying ANN to AA**

Of the three approaches described in Section 2 above, I will not consider the first approach further, because it treats an ANN just like another ordinary ML classifier. The second approach heavily pre-processes data to extract linguistic information from them, and then uses the information to improve the model which is developed by using the data as a whole, including the topic related content words. All models developed by the second approach have not been tested by using data obtained from a source different from the source of the training data or the test data. The third approach focuses on just designing a very sophisticated ANN and then letting it perform an end-to-end process. Both the second approach and the third approach require development of a complicated ANN and consume a lot of computing power. Those two approaches are not effective.

**3.2 My intended approach**

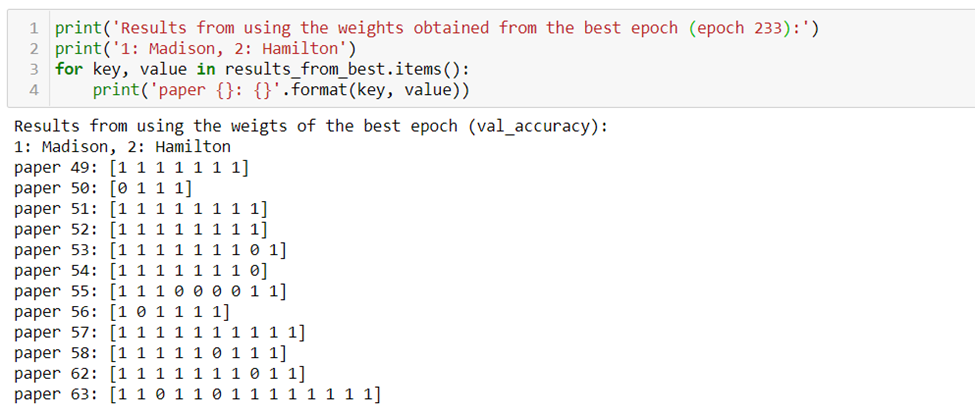
It is my experience that, to study human complied data, such as structured data (i.e., tables), using a simple ANN, such as a basic dense (fully connected) neural network (DNN) is sufficient (There are two examples described in Section 2.1. One is the model produced by the weerasinghe20 team and the other one was produced by me). On the other hand, to study data that are direct records of the nature, such as pictures and human speeches, one should opt for using a sophisticated network and should not pre-process the data heavily. A researcher on using ANN to study human speeches once told me, anecdotally, that the approach of first reducing human speeches to words or phonetic symbols and then feeding them to an ANN for analysis is always futile.

I found that written text is a kind of hybrid data. Written texts are records of naturally occurred matters (speeches) by using human invented symbols. Therefore, either too heavily pre-process text data (approach 2), or virtually not pre-processing text data and just feeding them to a very sophisticated ANN for it to learn from the data itself, will not be effective. Therefore, the aim of my research is to invent novel and effective ANNs and the corresponding data pre-procession methods, which are balanced between the two above-described approaches.

When doing the research, I will read most or all recent published papers about applying ANN to AA, learn cutting-edge ANN technologies and the mathematics behind them, learn and implement other partitioners’ code, and develop my computer programmes, to produce effective ANNs and their corresponding data pre-processing methods. Effective ANNs here means that the ANNs should not be too complicated and the data used should be pre-processed according to certain linguistic rules, but should not require too heavy pre-procession. An example of heavy pre-procession is forming a linearised parse tree for every sentence as described in Section 2.2.

**3.3 An example of using a simple ANN and a simple linguistic feature to perform AA**

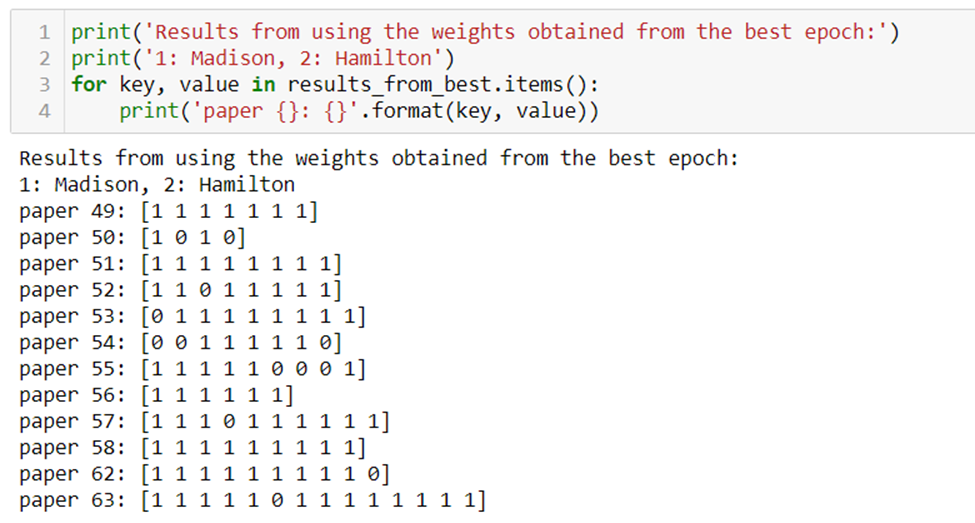
IN November last year (2020), I developed a simple ANN model that uses the 75 undisputed Federalist Papers as the training data to study who, Madison or Hamilton, wrote the 12 disputed papers (Please see Section 1). I extracted a simple linguistic feature from the data and used it to obtain the following result:



I was very excited because my finding is the same as that of Mosteller and Wallace:

On the basis of our data alone, Madison is extremely likely, in the sense of degree of belief, to have written all the disputed Federalists: Nos. 49 through 58 and 62 and 63, with the possible exception of No. 55. (Mosteller and Wallace, 1964, p. 263)

However, when I used another linguistic feature to perform the sample experiment, I got an unsatisfactory result. Therefore, I checked the code and the data. I then found that I had made a mistake when pre-processing the data. I had forgotten to delete the line feeders '\n' and the tab feeders '\t' from the data. I corrected the mistake and then applied the model to the two features again. The second feature still gave an unsatisfactory result. Very disappointed, I found the result produced by the first feature also deteriorated to the following:



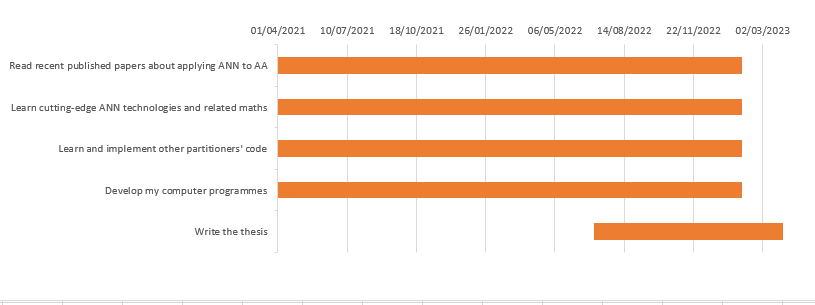
The model was in fact not able to single out paper No. 55 as the most undecisive paper and not able to determine who wrote paper No. 50, although it could still correctly attribute 11 out of the 12 disputed papers to Madison. I hope that I can improve the model by making it a bit more sophisticated in the near future.

**3.4 Computing required resources**

I found that a laptop can handle most common traditional ML projects. But a laptop cannot cope with common ANN projects. I need more computing power, especially relatively advanced GPUs and more RAM capacity.

**4. The Gantt chart**

The Gantt chart was produced based on the assumption that the research stage will start in the summer term 2021 and last for two years.



**References**

Bevendroff, J. et al. (2020) ‘Shared tasks on authorship analysis at PAN 2020’, in Jose J. et al. (ed.) *Advances in Information Retrieval, European Conference on Information Retrieval 2020*. Lecture Notes in Computer Science: vol. 12036, pp. 508-516.

Boenninghoff, B. et al. (2020) 'Deep Bayes factor scoring for authorship verification: notebook for PAN 2020 at CLEF 2020', *PAN at CLEF 2020*. Thessaloniki, Greece, 22-25 September. Available at: https://arxiv.org/abs/2008.10105 (Accessed: 12 December 2020).

Chiu, C. H. (2019) ‘Selecting, deriving and applying computational authorship attribution methods that are suitable for studying early modern dramas:

A preliminary exploratory systematic review’. *CS8002, ProfDoc Data Science*, University of East London. Unpublished. Available at: https://github.com/ericchchiu/CN8002\_and\_DS7001 (Accessed: 28 December 2020).

Chiu, C. H. (2020a) ‘A study about using machine learning classifiers KNN and SVM to perform authorship attribution analysis’. *DS7003, ProfDoc Data Science*, University of East London. Unpublished. Available at: https://github.com/ericchchiu/u1720146\_DS7003\_courseworkCodeAndData (Accessed: 28 December 2020).

Chiu, C. H. (2020b) ‘An assessment of two theses about authorship attribution’. Coursework for *DS7004, ProfDoc, Data Science*, University of East London. Unpublished. Available at: https://github.com/ericchchiu/u1720146\_DS7004\_courseworkCodeAndData (Accessed: 28 December 2020).

Cohen, E. (2017) ‘How to predict Quora question pairs using Siamese Manhattan LSTM’. *Medium*, 7 June. Available at: <https://medium.com/mlreview/implementing-malstm-on-kaggles-quora-question-pairs-competition-8b31b0b16a07> (Accessed: 28 December 2020).

Goldberg, Yoav (2017) *Neural network methods for natural language processing*, Williston, VT, USA: Morgan & Claypool.

Honnibal, M. (2016) ‘Embed, encode, attend, predict: The new deep learning formula for state-of-the-art NLP models’. Explosion, 10 November. Available at: <https://explosion.ai/blog/deep-learning-formula-nlp> (Accessed: 28 December 2020).

Jafariakinabad, F. and Hua, K. A. (2019) ‘Style-aware neural model with application in authorship attribution’, *8th IEEE International Conference on Machine Learning and Applications*. Boca Raton, Florida, 16-19 December 2019. Piscataway, New Jersey: IEEE: pp. 325-328.

Jafariakinabad, F. and Hua, K. A., (2020) ‘A self-supervised representation learning of sentence structure for authorship attribution’, Manuscript submitted to *Association for Computing Machinery*, Available at: https://arxiv.org/abs/2010.06786 (Accessed: 12 December 2020).

Juola, P. (2013a) ‘How a computer program helped show J K Rowling write a Cuckoo’s Calling: Author of the Harry Potter books has a distinct linguistic signature’. *Scientific American*, 23 August. Available at: https://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/#:~:text=By%20looking%20at%20Galbraith's%20language,author%20and%20discredit%20the%20tip (Accessed: 28 December 2020)

Juola, P. (2013b) ‘Rowling and "Galbraith": an authorial analysis’, *Language Log*, 16 July. Available at: https://languagelog.ldc.upenn.edu/nll/?p=5315 (Accessed: 28 December 2020)

Mosteller, F. and Wallace, D. L. (1964) *Inference and disputed authorship: The Federalist*. Reading, MA, USA: Addison-Welsey.

PyData (2017) *Applying the four step ‘Embed, Encode, Attend, Predict’ framework to predict document similarity*. 24 July. Available at: https://youtu.be/HfnjQIhQzME (Accessed: 28 December 2020).

PAN: Authorship verification 2020 (2020). Available at: https://pan.webis.de/clef20/pan20-web/author-identification.html#data (Accessed: 28 December 2020).

Rudman, J. (1998) ‘The state of authorship attribution studies: Some problems and solutions’, *Computers and the Humanities*, 31(4), pp. 351–365.

Rudman, J. (2010) ‘The State of NonTraditional Authorship Attribution Studies – 2010: Some Problems and Solutions’, *English Studies*, 93(3), pp. 259-274.

Sari, Y. et al. (2017) ‘Continuous N gram Representations for Authorship Attribution’, *15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain, 3-7 April 2017. Stroudsburg, Pennsylvania: Association for Computational Linguistics: Volume 2, Short Papers, pp. 267–273.

Weerasinghe, J. and Greenstadt, R. (2020) ‘Feature vector difference based neural network and logistic regression models for authorship verification: notebook for PAN at CLEF 2020’, *PAN at CLEF 2020*. Thessaloniki, Greece, 22-25 September. Available at: http://ceur-ws.org/Vol-2696/paper\_125.pdf (Accessed: 12 December 2020).

Zhang, R. et al., (2018) ‘Syntax Encoding with Application in Authorship Attribution’, *Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, 31 October- 4 November 2018. Stroudsburg, Pennsylvania: Association for Computational Linguistics: pp. 2742–753.

**Certificate of Completion of the research integrity module**:



