**ASSESSMENT AND INTERNAL VERIFICATION FRONT SHEET (Individual Criteria)**

**(Note : This version is to be used for an assignment brief issued to students via Classter)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Course  Title | Bachelor of Science (Hons). In Multimedia Software Development  Bachelor of Science (Hons). In Software Development  Bachelor of Science (Hons). In Networking | | | Lecturer Name & Surname | Gerard Said Pullicino | | |
| Unit Number & Title | | **ITPRJ-606-1606 IT Project** | |  |  | | |
| Assignment Number, Title / Type | | **Assignment 1** | |  |  | | |
| Date Set | | **25/03/2021** | **Deadline Date** | **24/05/2021** |  | | |
| Student  Name | **Mark Dagostino** | | **ID Number** | **349298M** |  | **Class /**  **Group** | **SWD 6.3A** |

|  |  |
| --- | --- |
| **Assessment Criteria** | **Maximum Mark** |
| • **AA 2 Implement a research structure for a given subject area** | 7 |
| • **AA 5 Organise collected data in a valid and correct way in own project.** | 7 |
| • **KU 3 Discuss the suitability of different research methodologies to different subject areas.** | 5 |
| • **AA 1 Select the correct research methodology for a specific subject area.** | 7 |
| • **KU 1 Recognize different research methodologies and their use within a project** | 5 |
| • **KU 4 Explain the different research methodologies used by different authors in a subject area** | 5 |
| • **KU 5 Explain the distinction between different research types for a specific topic** | 5 |
| • **KU 2 Identify areas of academic discourse within a specific subject area** | 5 |
| • **KU 6 Identify a valid number of research papers in a given subject area that are related to own project** | 5 |
| • **KU 7 Indicate how results may be presented in the context of own documentation** | 5 |
| • **AA 3 Assess collected information related to the implementation of the research process of own project** | 7 |
| • **AA 4 Present different research perspective in chosen subject area for a project** | 7 |
| • **SE 1 Appraise how results are collected by the majority of researchers in own chosen subject area** | 10 |
| • **SE 2 Create a valid set of results based on collected data for own project** | 10 |
| • **SE 3 Derive conclusions from valid scientific data correctly to determine results of own project** | 10 |
| **Total Mark** | 100 |

|  |
| --- |
| **Notes to Students:** |
| * This assignment brief has been approved and released by the Internal Verifier through Classter.      * Assessment marks and feedback by the lecturer will be available online via Classter ([Http://mcast.classter.com)](http://mcast.classter.com/) following release by the Internal Verifier      * Students submitting their assignment on Moodle/Unicheck will be requested to confirm online the following statements:     **Student’s declaration prior to handing-in of assignment**   * + I certify that the work submitted for this assignment is my own and that I have read and understood the respective   Plagiarism Policy    **Student’s declaration on assessment special arrangements**   * + I certify that adequate support was given to me during the assignment through the Institute and/or the Inclusive Education Unit.   + I declare that I refused the special support offered by the Institute. |

# ICT Project Assignment

This assignment contains tasks which will give you practical skills in the preparation of your dissertation. Completing these tasks will give you practice in the skills that you will need to use to prepare the supporting documentation and presentation of your dissertation. You are therefore encouraged to perform ALL of the tasks below as part of your submission for this subject. Please note that this assignment is assessed *independently* of the dissertation and carries its own assessment, which is 100% of the marks for the ICT project module in the third year of the degree.

Your submission for this course will be a git enabled folder with a minimum of one commit per task. For each commit description, please include the location of the files and all the information required to access the files necessary for each task. The upload will be to a unicheck enabled submission on Moodle at the end of the course.

Unless advised otherwise, documents submitted for this assignment should either be PDF files or HTML web pages which can be accessed offline with the appropriate relative paths.

## Task 1 - Concept mapping

You are required to use Xmind to create a valid concept map of either your own dissertation, or a dissertation of your choice from the past dissertations repository.

Your concept map should follow the following rules:

1. Leaf nodes should be sources supporting a concept.
2. No two identical leaf nodes (similar elements should be linked by relationships).
3. All relationships should be labelled.
4. All referenced papers should be included in the concept map.
5. All referenced papers should be listed in a separate bibliography in correct format.

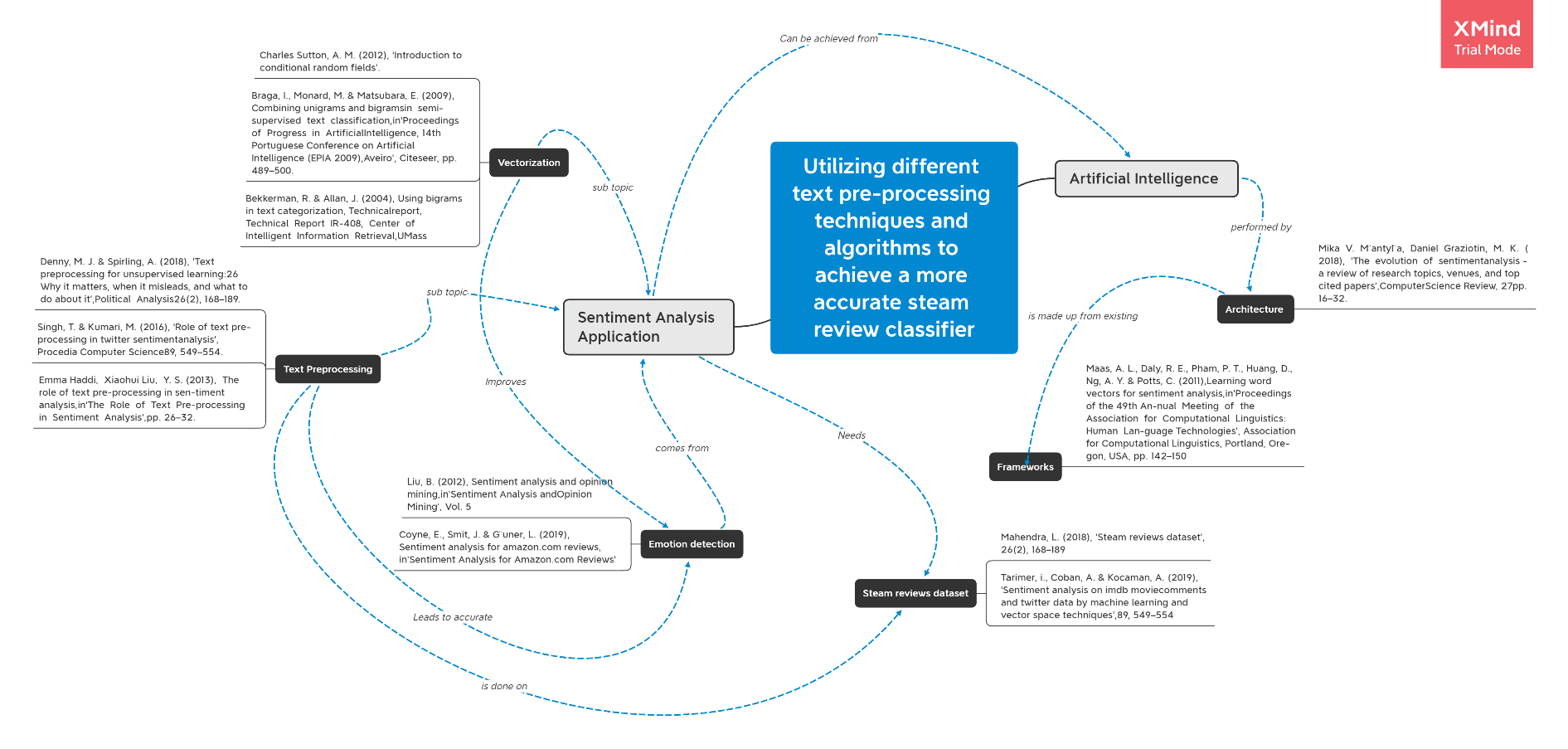
The concept map should be uploaded in .png format in a resolution of at least 1920x1080 (ensure that the assessor can easily zoom in and read the text of the citation).

Apart from your concept map include a 150 word paragraph explaining the limitations of your concept map, and any observations you may have on the limitations of the research subjects that have been explored. These may be justified if you are referring to your own dissertation. Please note that this is an assessment of the concept map NOT your literature review. (14 marks)

### Criteria Assessed

* AA 2 Implement a research structure for a given subject area
* AA 5 Organise collected data in a valid and correct way in own project.

|  |  |  |
| --- | --- | --- |
| **Poor quality work (0-5 marks)** | **Moderate quality work (6-11 marks)** | **Top quality work (11-14 marks)** |
| No clear structure defined, no understanding of the difference between sub topics and relationships, few to no sources referenced | Clear structure, good understanding of topics and subtopics, some relationships missing, some sources within the papers omitted | Clear structure, good understanding of topics and subtopics, all sources included and correctly related and  referenced |



**Bibliography**

Charles Sutton, A. M. (2012), ‘Introduction to conditional random fields’.

Braga, I., Monard, M. & Matsubara, E. (2009), Combining unigrams and bigramsin semi-supervised text classification,in‘Proceedings of Progress in ArtificialIntelligence, 14th Portuguese Conference on Artificial Intelligence (EPIA 2009),Aveiro’, Citeseer, pp. 489–500.

Bekkerman, R. & Allan, J. (2004), Using bigrams in text categorization, Technicalreport, Technical Report IR-408, Center of Intelligent Information Retrieval,UMass

Denny, M. J. & Spirling, A. (2018), ‘Text preprocessing for unsupervised learning:26

Why it matters, when it misleads, and what to do about it’,Political Analysis26(2), 168–189.

Singh, T. & Kumari, M. (2016), ‘Role of text pre-processing in twitter sentimentanalysis’,Procedia Computer Science89, 549–554.

Emma Haddi, Xiaohui Liu, Y. S. (2013), The role of text pre-processing in sen-timent analysis,in‘The Role of Text Pre-processing in Sentiment Analysis’,pp. 26–32.

Liu, B. (2012), Sentiment analysis and opinion mining,in‘Sentiment Analysis andOpinion Mining’, Vol. 5

Coyne, E., Smit, J. & G ̈uner, L. (2019), Sentiment analysis for amazon.com reviews,in‘Sentiment Analysis for Amazon.com Reviews’

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y. & Potts, C. (2011),Learning word vectors for sentiment analysis,in‘Proceedings of the 49th An-nual Meeting of the Association for Computational Linguistics: Human Lan-guage Technologies’, Association for Computational Linguistics, Portland, Ore-gon, USA, pp. 142–150

Mahendra, L. (2018), ‘Steam reviews dataset’,26(2), 168–189

Tarimer, i., Coban, A. & Kocaman, A. (2019), ‘Sentiment analysis on imdb moviecomments and twitter data by machine learning and vector space techniques’,89, 549–554

Mika V. M ̈antyl ̈a, Daniel Graziotin, M. K. (2018), ‘The evolution of sentimentanalysis - a review of research topics, venues, and top cited papers’,ComputerScience Review, 27pp. 16–32.

Limitations:

Since the dissertation related to this concept map focuses whether or not a classifier is able to improve from the use of different text pre-processing techniques, the literature review heavily relies on the research of text pre-processing and vectorization techniques. This may take away from the exploration of different datasets, different algorithms which may also influence the result of the classifier. One limitation I have noticed during the research of the literature review is that, unfortunately, there is a lot of research which contradicts the other when it comes to how a classifier detects text. For example, some research state that certain stop words, may help the classifiers final sentiment, when others say that it does not influence the final predication at all. This leads me to believe that their research/classifier relies solely on one type of structure (Dataset & algorithm). Since I will need to explore multiple structures, it may take away from the consistency and efficiency of the classifier.

When you write your methodology, a specific order of elements needs to be observed. For task 2, you are required to write an explanation of the elements needed in your methodology and the reason why each section is required. Aim for a wordcount of approximately 1,200 words for this section, which should contain the following three elements:

Ethics:

Explain what mitigations are required to ensure ethics are correctly followed at the following stages of research (Refer to Cresswell):

1. Prior to beginning the study
2. Beginning the study
3. Collecting data
4. Analysing data
5. Reporting, Sharing and Storing data

Introduction:

List what elements *have* to be included in the introduction to any methodology.

Research Design:

Justify a research design for a specific dataset (you may either use your own dataset or the dataset of research that was used in a past dissertation). Explain why you have chosen that specific research design from the options of Quantitative, Qualitative or Mixed Methods. One example of a justification is a *deficiency in past research*. You need to at least define that justification and add another two justifications to your own research approach based on your chosen data.

Ethics:

Cresswell states that prior to commencing the study, the researcher needs to consider the code of ethics. This is why at the beginning of any research study, several considerations need to be taken. In my case, a proposal was created. The proposal includes the hypothesis and research questions of the research. The type of the study, such as quantitative, qualitative or mixed, a mock-up of the literature review, a brief overview of the research design which includes the methodology and arguably one of the most important sections, any potential ethical issues which can arise at any point of the research study. After the proposal is approved, the research study may start.

Since in my dissertation we are focusing on the sentiment of text, those of which are being written by consumers and are usually directed towards a company or their relating property. There were several ethical considerations that were taken in order to keep both parties involved (Myself and the experimental group) in a beneficial state. Firstly, the study uses two different datasets, both of which are completely public and available on the company’s official platform for research purposes and all relating articles were cited according to the platform’s requirements.

Creswell refers to analysing data as follows: When analysing the data, the researcher must avoid going native. In quantitative research, this means ignoring data that confirms or otherwise rejects any personal hypothesis that the researcher may have. An important note they also mention is to avoid disclosing only positive results. This was found apparent when conducting our literature review, this is because It is misleading to withhold important findings or to cast the findings in a positive light to the participants or researchers’ inclinations. Finally, another very important note is to respect the privacy of participants.

When analysing our datasets, none of the provided data were manipulated in any form, however, despite being public some columns were omitted to make sure any possible exploitation does not occur in the interest of both parties. This includes any public names attached to the reviews, dates, any partner companies and/or any sensitive data which may affect a consumer or an organisation. The only data needed for our research was a real review, and its label, being either positive or negative. This was done simply to stay lenient in our research studies goal and in the interest of relating parties. Of course, this has its disadvantages, for example, certain statistics may not be possible, however it is a necessary step towards a fair research study.

For evidence, data findings, or conclusions, Cresswell states that researchers need to provide a precise account of the data. This may require interviews for the researcher and participants. Plagiarism is prohibited and researchers should always give recognition for the work of others. Quotation’s mark should always show the exact words taken from others. Some participants may also want their identity to remain confidential and this should always be respected. When reporting results, the communication must be clear and straightforward. Language must not include words against any individual or organisation, sexual orientation, or disability. Once the data is analysed, it is important that is kept for a reasonable period.

Another important part of our research is how the results were displayed. Since the main objective is to find whether text pre-processing techniques is able to improve a model’s overall performance, only relating statistics were shown, such as a model’s accuracy. No correct or incorrect reviews were shown in the dissertation in order to stay neutral. Finally, when it comes to sharing or storing any results or data, Only the data retrieved from the official resource is stored online, the results from the data are only written in the dissertation itself and models have to be rerun every time results wish to be retrieved, possibly meaning similar or different results may be returned.

Introduction:

1. The research problem must be discussed.
2. A review of past studies about the problem we are trying to solve.
3. It is important to mention the purpose statement.
4. List all independent and dependant variables.
5. The significance of the study in relation to a company’s interest.
6. To identify any issues that may occur when collecting the data that is going to be used.

Research design:

The research design relating to my dissertation is a quantitative research. This is simply because the data being used will be returning numbers and statistics. No Surveys or Questionnaires will be conducted for the purpose of this research.

In order to remove any discrepancies in the results of the data, and so that we are able to compare accurately with previous research, an average was taken from all the models created. Each model was ran at least 5 times. The results listed are the accuracy, Recall, Precision and F1 Score of each model.

Most research studies found used only a number of techniques and only tested their models out with and without said method and compared. I believe there is a lot more potential in this research design which is why all text pre-processing techniques documented in the literature review was included and tested.

In order to keep also what is stated above in ethics in check, going native is wrong, this is why no correct or incorrect labels were shown throughout the research. In our prototype we can see what the models predicted; however, we did not document this to respect both parties. Only the performance was documented. This was chosen after analysing a research paper which focused heavily on only correctly labelled reviews.

(12 marks)

### Criteria Assessed

* KU 3 Discuss the suitability of different research methodologies to different subject areas.
* AA 1 Select the correct research methodology for a specific subject area.

|  |  |  |
| --- | --- | --- |
| **Poor quality work (0-4 marks)** | **Moderate quality work (5-8 marks)** | **Top quality work (8-12 marks)** |
| No mitigations mentioned, poor understanding of the reasons for ethics provisions, poor understanding of the required recordkeeping for good research. Justifications for research do not make sense. | Some mitigations mentioned, the phases are tackled correctly and references to Cresswell are made correctly. One justification for the research being carried out mentioned and explained | All ethical issues and their mitigations at the different phases of the study are mentioned in detail. Three justifications mentioned in detail and justified logically and convincingly. |

## Task 3

In this section, please find the following sample purpose statements (Cresswell p.170):

1. Case study/Qualitative study The purpose of this \_\_\_\_\_\_ (strategy of inquiry, such as ethnography, case study, or other type) study is (was? will be?) to \_\_\_\_\_\_ (understand? explore? develop?

discover?) the \_\_\_\_\_\_ (central phenomenon being studied) for \_\_\_\_\_\_ (the participants, such as the individual, groups, organization) at \_\_\_\_\_\_ (research site). At this stage in the research, the \_\_\_\_\_\_ (central phenomenon being studied) will be generally defined as \_\_\_\_\_\_ (provide a general definition).

1. Quantitative study The purpose of this \_\_\_\_\_\_ (experiment? survey?) study is (was? will be?) to test the theory of \_\_\_\_\_\_ that \_\_\_\_\_\_ (describes outcomes) or (compares? relates?) the \_\_\_\_\_\_ (independent variable) to \_\_\_\_\_\_ (dependent variable), controlling for \_\_\_\_\_\_ (control variables) for \_\_\_\_\_\_ (participants) at \_\_\_\_\_\_ (the research site). The independent variable(s) \_\_\_\_\_\_ will be defined as \_\_\_\_\_\_ (provide a definition). The dependent variable(s) will be defined as \_\_\_\_\_\_ (provide a definition), and the control and intervening variable(s), \_\_\_\_\_\_, (identify the control and intervening variables) will be defined as \_\_\_\_\_\_ (provide a definition).

Select three academic papers you are referencing in your literature review (or select three sources from a past dissertation of your choice) and create a purpose statement for the academic papers being referenced. You need to include 3 purpose statements for 3 different sources, and correctly cite the sources you are using. Select the correct script and try to include all the elements of the scripts. Cresswell also includes a reference to a mixed methods purpose statement, if the study of your choice is a Mixed methods study.

1. The purpose of this research was to test if sentiment analysis techniques are also feasible for the application of real product reviews. The main resource used in this study is the dataset of amazon reviews made up of 4 million reviews under different categories. The independent variables were defined as three different algorithms which were all fitted with an equal test/train ratio. This would ultimately return the dependent variables defined as accuracy, precision, recall and f-score.

Coyne, E., Smit, J. & G ̈uner, L. (2019), Sentiment analysis for amazon.com reviews,in‘Sentiment Analysis for Amazon.com Reviews’

1. The purpose of this experiment was to put the role of text pre-processing to the test. The independent variables were defined as three different feature extraction techniques, tested using one algorithm and on one dataset. The dependant variables were shown as accuracy, precision, recall and f-score.

Emma Haddi, Xiaohui Liu, Y. S. (2013), The role of text pre-processing in sen-timent analysis,in‘The Role of Text Pre-processing in Sentiment Analysis’,pp. 26–32.

1. The purpose of this case study was to explore feature extraction combinations for a semi-supervised neural network. 5 text datasets were gathered and were used to run self-training and co-training simulations. These independent variables are being extended to include even more datasets, as well as to include text representations. The dependable variable is measured by the percentage error and the accuracy.

Braga, I., Monard, M. & Matsubara, E. (2009), Combining unigrams and bigramsin semi-supervised text classification,in‘Proceedings of Progress in ArtificialIntelligence, 14th Portuguese Conference on Artificial Intelligence (EPIA 2009),Aveiro’, Citeseer, pp. 489–500.

(15 marks)

### Criteria Assessed

* KU 1 Recognize different research methodologies and their use within a project
* KU 4 Explain the different research methodologies used by different authors in a subject area
* KU 5 Explain the distinction between different research types for a specific topic

|  |  |  |
| --- | --- | --- |
| **Poor quality work (0-5 marks)** | **Moderate quality work (6-10 marks)** | **Top quality work (11-15 marks)** |
| Wrong purpose statement template chosen for a paper. The purpose statement does not match the research type. Only one purpose statement included | Three purpose statements mentioned and cited correctly. In some cases, the wrong purpose statement template was used, but the study was correctly explained. | Three purpose statements are mentioned and cited correctly based on the correct research papers and are the correct purpose statements for the research model being used.. |

Now it is time for your own opinion. Do you agree with the conclusions of the sources you created the purpose statements for? If you agree, explain why you agree and cite elements highlighting your agreement in the dissertation. If you disagree, explain why you disagree and cite elements highlighting your disagreement. You need to make at least 5 points of agreement and disagreement in this section. Are any of these papers being used in your own dissertation?

Since my dissertation is based on the research conducted by all these individuals, I wanted to test out their theories myself on relatively similar data.

1. Despite showing decent results throughout, I do believe having a dataset made up of 4 million mixed reviews can be, not only confusing but also time-consuming in terms of text preparation, especially if you have multiple techniques being tested.
2. Their research design however, followed a great and solid structure, making use of different independent variables to gather as much information and ultimately build a strong classifier. They also utilized most dependable variables to make sure results are returned correctly and do not show inaccurate results.

Coyne, E., Smit, J. & G ̈uner, L. (2019), Sentiment analysis for amazon.com reviews,in‘Sentiment Analysis for Amazon.com Reviews’

1. Since my dissertation compares a lot with Haddi’s work, there is a lot to naturally agree with, an infamous example would be that the omission of certain words, such as “if” or “a”, would not generally change a sentence meaning or sentiment, but in contrast, can actually improve a classifiers performance. This research focuses a lot on the role of text pre-processing, as does mine, and the methods used are some of the most popular methods to date, which are known to give good results on many different datasets. In terms of their hypothesis, general research, and research design, I agree with what they state and practice all throughout.

Emma Haddi, Xiaohui Liu, Y. S. (2013), The role of text pre-processing in sen-timent analysis,in‘The Role of Text Pre-processing in Sentiment Analysis’,pp. 26–32.

1. Although this research did not focus on “The best performance” I do not agree with how the results are being displayed. Visual representations focus heavily on correctly labelled results which may be misleading, and the research only shows the accuracy of each method. This may be very misleading, because a classifier can have a high accuracy but a very low F1-score.
2. Despite my disagreement, their research is very interesting and displays a very interesting research question. being partially used also in my own dissertation, I do agree with a big part of their research design, testing their classifiers on multiple datasets and comparing the results, comparing the individual methods with both combinations and despite having a relatively long training time, also testing taking it a step further briefly including tri-grams to display interesting research for future candidates.

Braga, I., Monard, M. & Matsubara, E. (2009), Combining unigrams and bigramsin semi-supervised text classification,in‘Proceedings of Progress in ArtificialIntelligence, 14th Portuguese Conference on Artificial Intelligence (EPIA 2009),Aveiro’, Citeseer, pp. 489–500.

(10 marks)

### Criteria Assessed

* KU 2 Identify areas of academic discourse within a specific subject area
* KU 6 Identify a valid number of research papers in a given subject area that are related to own project

|  |  |  |
| --- | --- | --- |
| **Poor quality work (0-3 marks)** | **Moderate quality work (4-6 marks)** | **Top quality work (7-10 marks)** |
| No logical conclusions drawn, no understanding of agreeing or disagreeing in an informed way with the cited sources. | Some logical agreements and disagreements shown. Minimum 3 points of agreement and  disagreement | Three purpose statements are mentioned and cited correctly based on the correct research papers and are the correct purpose statements for the research model being used in the papers is referenced. |

Explain the meaning of the following terms, with at least one example based on your own observations for each:

1. **Experimental protocol** – An experimental protocol, I believe is simply the hypothesis we are trying to test in our research along with the research questions we are trying to answer. This can be answered with a set of steps that we create or choose to follow. For example, what will be the process of creating a sentiment review classifier, how long will the test last. Etc.
2. **Independent variables** – an independent variable is an “object” that is fully independent of any other factor. For example, a movies age won’t change based on how many movie reviews they have. An algorithm is also considered an independent variable.
3. **Dependent variables** – a dependent variable, on the other hand, is an “object” that depends on other factors, for example, the classifiers accuracy depends on the training dataset given.
4. **Experimental design** – Experimental design I believe is how different sections/people are assigned to an experiment. For example, when conducting a survey, we may assign different people to answer different research questions.
5. **External validity of an experiment** – External validity I believe is the feedback given to by a third party of one’s research. Possibly someone who is able to validate if the work created is valid. For example, our work will be externally tested by third party sources and given the appropriate feedback based on results and academic research.
6. **Sampling types** – There exists different types of sampling methods. I believe sampling has to do with a group of objects or people that are taken from a larger amount of population. For example a poll.
7. **Coding** – Coding I believe is fairly straight forward. It is the bulk of code that makes up our experimental project. For example, the code that makes up a sentiment review classifier. This includes, the dataset being used, cleaned, and tested.
8. **Research triangulation** – Triangulation is when we mix different research methods. In my scenario, I am mixing different techniques to see whether it would affect a classifier in a positive manner

(19 marks)

### Criteria Assessed

* KU 7 Indicate how results may be presented in the context of own documentation
* AA 3 Assess collected information related to the implementation of the research process of own project
* AA 4 Present different research perspective in chosen subject area for a project

|  |  |  |
| --- | --- | --- |
| **Poor quality work (0-6 marks)** | **Moderate quality work (7-13 marks)** | **Top quality work (14-19 marks)** |
| Incorrect definitions for any of the terms above. No examples given. | Correct definitions for the above terms given, however the examples given are either the typical examples found in the actual reference of the term or the examples are not explained. | Correct definitions given, with examples taken from samples of the student’s own research, and explained using examples given by the student himself/herself |

Critically analyse the process of the development of this report. Write a section outlining how your understanding of your subject matter has changed in the context of the process of doing academic research.

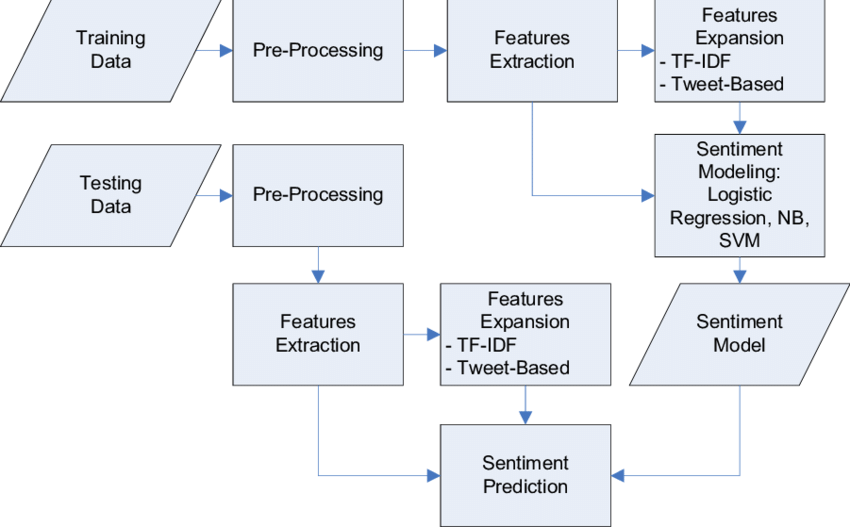
This section needs to contain an element of self-reflection, and there is the requirement of analysis on the experience of the change in understanding of the subject that you are exploring in your dissertation or a past dissertation of your choice.

Please use the following points as guidance to answer this question:

1. What was your starting understanding of the problem area?
2. What was the initial choice of method with respect to methodology?
3. How did the methodology change based on the literature review?
4. How did the chosen methodology affect the presentation of results?
5. What are the limitations highlighted in your or other dissertations?
6. What possible areas of related study do you believe are feasible, based on the presented results?

I initially chose sentiment analysis, because the idea of a model that I helped create and develop, predicting a text’s sentiment, sound quite impressive at the time. I got into it learning a simple pipeline of how it should work. Something along the lines of get your data, clean out any noise that might be in it, feed a portion of it to a model and test your model and receive results. I soon came to realize how much more you are able to do with the same methods, Firstly, it’s not only text that you’re able to base models on, audio, video, images all have their own research dedicated towards them and are very interesting.

The way I wanted to work was simply based on other people’s research, I wanted to see what people were using, combine their methods, and see if I can create something that performs even better, so my methodology was heavily based on what research was out there. When conducting research however, it got even harder to choose what method I was focusing on, because even more topics came into play. Such examples would be supervised research vs unsupervised research or if the dataset were compatible with the independent variables etc. I soon realized that you can’t grab what people are using and compare them directly with other methods because every research design is different. A simple pipeline very soon started to build into quite a complex one.



I still managed to compare different methods, however. The independent variables that I chose, were based on the most popular methods being used. I chose two subjects which I enjoyed, movies and games and found a review dataset for each. All were labelled and needed supervised learning algorithms to compare correctly. I also chose the most popular text pre-processing methods based on what other people reported, so that we can indeed compare the results. Finally, I also chose two very popular algorithms and compared my own models and saw how the results differed from dataset to dataset.

In total, I gathered 26 different models, 13 for one dataset, and 13 for the other. All using both popular and new methods of research and all had extremely interesting results. Results which were reported to be very good, performed worse on mine, Text cleaning techniques which claimed to rule out characters or words which have no sentiment, such as “😊” or “☹” indeed showed some sentiment. Displaying these results was quite a challenge because I felt as if it was hard to dismiss or disagree with someone’s research. But it all went down to the dataset, how the model trained, and how it was tested.

Not all these steps could be replicated, and in my opinion, it also shouldn’t, since we’re all exploring different areas within the same field. And that made it even more interesting. All models performed relatively decently on all dataset, which showed that models could indeed use different datasets and techniques and perform even better, which was my whole hypothesis to begin with. I believe I did cover what my research questions and hypothesis were asking, however that’s not to say that the research didn’t have limitations, to begin with, both datasets were rather large in their own respect, and therefore some feature extraction techniques, such as tri-grams had to be oppressed from the research, simply because the matrix created proved to be too large and training would’ve taken a lot longer than it’s relating methods. A research going into trigrams and how it can show further improvement would be quite an interesting read. In my research only two algorithms were created, SVM and LR. An extended research on Naïve Bayes would also be quite interesting to compare.

One very interesting yet challenging study I would like to see, is sarcasm detection. While grabbing sentiment from text, you can never be quite sure if what a person is writing is true or not. And whilst most of the time, the emotion of a review is clear for a human, it is not always the case, imagine what it would be like for a model. A research like this, I believe could be very beneficial to social media and even analytic companies. Another research would I believe would prove to be challenging is double negative or double positive detection. This would be a research study were you would differentiate between a phrase such as “the car is not bad” and “your driving skills are not too bad”. The first phrase is actually a positive sentiment and the second one is just slightly more negative.

(30 marks)

### Criteria Assessed

* SE 1 Appraise how results are collected by the majority of researchers in own chosen subject area
* SE 2 Create a valid set of results based on collected data for own project
* SE 3 Derive conclusions from valid scientific data correctly to determine results of own project

|  |  |  |
| --- | --- | --- |
| **Poor quality work (0-10 marks)** | **Moderate quality work (1120 marks)** | **Top quality work (21-30 marks)** |
| No thought given to understanding the process of the dissertation. Poor explanation of the strength of the process and how it changes the understanding of a subject. | Fair understanding of the academic process, with a good knowledge of the process involved. Examples given from own experience, illustrated by explanations from own or others literature | A clear understanding of the academic process. Explanation illustrated by examples from own or other sample dissertation, showing the way that the dissertation is structured and how the dissertation was built section by section. |