# Detailed Project Proposal

|  |  |
| --- | --- |
| First Name: | Jess |
| Last Name: | McGowan |
| Student Number: | 1703725 |
| Supervisor: | John Isaacs |

## Defining your Project

**1.1 Project title**

**Help:** a brief statement about what you are actually going to do.

|  |
| --- |
| A machine learning based analysis of social media sentiment, and the impact of mainstream media. |

**1.2 Background**

**Help:** Provide the background to your project. This section should highlight the main topics in the area you are going to research. Essentially what is the project about, what has been done before and why is this project important? ~500 words

|  |
| --- |
| This project is about performing sentiment analysis on tweets about Brexit, and creating a machine learning algorithm that can categorise tweets based on whether they are positive, neutral or negative in terms of sentiment displayed, and whether this changes over time in a significant way. I picked Brexit as a subject due to its timeliness and the controversy – there should be a lower number of neutral sentiment tweets compared to the number of positive and negative tweets.  Another aspect of this project will be investigating whether sentiment changes are triggered by particular news articles being published by mainstream outlets, and if the prevailing Twitter sentiment shifts before publication (e.g. that articles have been written based on a viral Twitter story), or if the change comes after publication (e.g. that an article has provoked an emotive response on Twitter).  Twitter sentiment analysis has been researched before, on a number of different topics. An early study by Go et al from Stanford University in 2009 explains some of the models that were used for their analysis of a set of tweets collected by searching for a number of subjects: Naive Bayes and Maximum Entropy, in which they achieved an accuracy rate of 83% using their models. As this study was carried about before Twitter introduced support for emojis, newer techniques will have to be used to properly account for emojis that convey sentiment.  Another study has been carried out by Agarwal et al, from Columbia University in 2011, on a set of tweets with no specific constraints on tweets (including selecting tweets from all languages, and using Google translate to convert the tweets into English. The techniques used in this study included an emoticon dictionary which gave each version of emoticons a polarity indicator to indicate sentiment, as well as expanding common web-based acronyms such as lol and gr8 to ‘laughing out loud’ and ‘great’ respectively.  For sentiment analysis on a specific topic, a 2015 study by Palomino et al, performed sentiment analysis on tweets relating to “Nature-deficit disorder” (NDD). These tweets were collected from a number of hashtags that had some relation to NDD, whether it is one that explicitly refers to NDD, such as #naturedeficitdisorder, or one that refers to the concepts, such as #outdoors. The researchers used the AlchemyAPI to perform the sentiment analysis of the collection of tweets, based on whether the tweet is positive, negative or neutral, and to what level the sentiment is expressed (whether the sentiment is strong or weaker). The analysis showed the sentiment analysis for each hashtag used in this study, as well as overall sentiment trends regarding the tweets relating to NDD.  In terms of Brexit-related studies, Landsdall-Welfare, Dzogang and Christianini (2017) from University of Bristol carried out Twitter based sentiment analysis around the time of the Brexit referendum, measuring levels of positivity, negativity, anger, anxiety and sadness in tweets posted from the UK, without specifying hashtags or keywords.  Sources:  Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R., 2011, June. Sentiment analysis of twitter data. In Proceedings of the workshop on languages in social media (pp. 30-38). Association for Computational Linguistics.  Go, A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12).  Lansdall-Welfare, T., Dzogang, F. and Cristianini, N., 2016, December. Change-point analysis of the public mood in UK twitter during the brexit referendum. In Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on (pp. 434-439). IEEE.  Palomino, M., Taylor, T., Goker, A., Isaacs, J. and Warber, S. 2016. The online dissemination of nature-health concepts: lessons from sentiment analysis of social media relating to "Nature-Deficit Disorder". International journal of environmental research and public health [online], 13(1), article number 142. Available from: https://doi.org/10.3390/ijerph13010142. |

**1.3 Motivation**

**Help:** To whom is this project important? A project must address a question/problem that generates a small piece of new knowledge/solution. This new knowledge/solution must be important to a named group or to a specific client (such as a company, an academic audience, policy makers, people with disabilities) to make it worthwhile carrying out. This is the ***motivation*** for your project. In this section you should address who will benefit from your findings and how they will benefit. ~300 words

**Example** 1: If you intend to demonstrate that a mobile application that automates class registers at RGU will be more efficient than paper-based registers - the group who would be interested in knowing/applying these findings would be both academic and administrative staff at RGU and they would benefit by time saved and a reduction in their administrative workload.

**Example** 2: You are demonstrating that a particular 3D model design increases realism in 3D environments. The group that would be interested would be games designers or developers of 3D virtual environment applications. They would benefit from producing more realistic environments that could increase sales of their products.

**Example** 3: You have designed a new network topology for IrishOil plc’s new Aberdeen headquarters. The interested group would clearly be IrishOil. They would benefit from easier maintenance and improved security of their computer network.

|  |
| --- |
| The main groups that would be interested in this algorithm will be politicians, journalists, and marketing analysts.  Journalists can see the effect that the media has on public perception, as this project will be investigating to see if there is a correlation in a change of public sentiment on Brexit based on the main news articles about Brexit, and if certain news media sources provoke a greater reaction (either positive or negative) than others.  Politicians can see what public sentiment is regarding Brexit in particular, and for some prominent politicians who appeared frequently in news articles, they can see the direct impact that they have had on the general sentiment of Twitter (i.e. if an headline has a direct quote which triggers a shift in the overall sentiment).  Marketing analysts can use the basis of the algorithm to see public perception of companies that they are employed by, and can alter their business and marketing strategies based on the findings – such as if a negative Twitter sentiment spike follows a corporate announcement, or if a positive spike occurs after a tweet goes viral showing the company in a positive manner. |

**1.4 Aim & Objectives**

**Help:** Outline what are the main things your project is going to do and what steps or milestones will be used to achieve this aim. The Aim is unlikely to change throughout your project; however, the objectives are likely to adapt to your ongoing research and development. In particular it is highly likely that you may wish to split objectives into sub-objectives as work progresses. A good clear set of objectives give you something to evaluate your final project against.

**Example** : For the timetable app outlined above

Aim: To create a functioning attendance application that efficiently automates the taking of class registers.

Objective 1: study existing register system in place at RGU and identify weaknesses

Objective 2: research existing automation technology’s and identify and evaluate those that may be appropriate to taking in class registers

Objective 3: Implement chosen technologies to create prototype application

Objective 4: Conduct user trials to evaluate capabilities of prototype application

Objective 5: Create a refined application incorporating feedback from user trials

|  |
| --- |
| **Aim:** To create an algorithm that can successfully predict whether a Brexit-related tweet is positive or negative in its sentiment, and to see how this sentiment changes over time.  **Objective 1:** Pre-process text-based data for creating a machine learning algorithm  **Objective 2:** Research different libraries for sentiment analysis  **Objective 3:** Create algorithm using different models, and compare the effectiveness of each one.  **Objective 4:** Using data visualisation techniques to see how Brexit sentiment on Twitter changes over time  **Objective 5:** Investigate whether there is correlation between news articles and sentiment changes. |

**1.5 Key Techniques**

**Help:** Perform some initial research into the area and outline what techniques you my research in further detail here. The techniques you cover here should include references to the papers where you have sourced the information. The techniques mentioned here are very likely to become the section headers in your literature review.

|  |
| --- |
| Data visualisation techniques – how to display, and how to interpret findings.  Using software libraries to conduct sentiment analysis on tweets, and evaluating different libraries.  Develop a machine learning algorithm to conduct sentiment analysis automatically on tweets.  Social understanding – how Twitter works, herd mentality in the context of viral tweets, reactionary posts. |

**1.6 Legal, Social, Ethical, Professional and Security issues**

**Help:** Here you should discuss any legal, social, profession and security issues that you believe may occur during the course of your project. It is not acceptable to write none in this box, all projects, regardless of focus will have to address issues in one, or more, of these categories. This is an extremely important part of your honours project to which there is no correct answer, this section must be fully discussed with your Honours Supervisor.

**Example 1** : In the class register example above – there would be a Legal and Security issue with the gathering and storage of student data. There may be a social constraint as you may be relying on a user to have access to a specific technology. There will need to be consideration of user accessibility.

**Example 2** : A 3D model design may have ethical considerations in its evaluation. What if your model made users feel nauseous. Social constrains may again be access to technology or accessibility issues.

**Example 3** : You network design need to adhere to specific company policies. You would need to consider the possibility that your design could be wrong, compromising the company’s security.

|  |
| --- |
| This project will use user-created content in the form of Twitter posts for analysis, which are publically available on Twitter. Personal information such as display names, account IDs, location and profile information will be removed from the data set before processing. As personal information will not be stored, data protection guidelines such as GDPR will not apply. The Twitter terms of service state that publically posted tweets are available for processing by any individuals or organisations, as they hold the licence for any content published.  The news articles used in this project will contain personal information such as names of specific people, some of which will be public figures in politics, and others will be private individuals who have been quoted. This will need to be accounted for when analysing the effect that certain news stories have on the general sentiment of Twitter users. |

**1.7 Project Plan**

**Help:** This is the project plan as to how you will go about achieving the objectives of the project.

**Example**: In the class register example above the research plan may involve:

Collecting and analysing paper-based registers in a given class on five occasions.

Identifying the error rate average on these occasions

Researching existing automation techniques

Designing and implementing a mobile application that automatically records attendance in class.

Deploying the application in the class on five occasions.

Identifying the error rate average of the mobile application on these occasions.

Comparison of data and summary of findings.

|  |
| --- |
| 1. Collecting tweets about Brexit, and processing the data before analysis such as removing personal information from the posts, and modifying the language and emojis used in tweets for text based analysis. 2. Researching existing libraries for sentiment analysis, and evaluating their effectiveness. 3. Researching different models to use for machine learning algorithms. 4. Create an algorithm to perform sentiment analysis on the data. 5. Visualise the findings to show how sentiment changes over time. 6. Examine if extreme points on the graph have a correlation to the publication of news articles about Brexit, and what the effect the article had. 7. Evaluate findings, and whether certain media companies had a greater impact on Twitter sentiment. 8. Explore further uses of the algorithm outside of Brexit-related sentiment.   See Appendix 1 for the initial project plan. |

**1.8 Ethics Form**

**You must include in your signed ethics form in this submission or you will not be able to continue the project.**

**Appendix 1 – Initial Project Plan**

