Sentiment Analysis

October 2019

By

Brian Davis

Word count: 2899

Contents

1. Project background and purpose

1.1. Objectives

With the emergence of Social Media as an area where most people go to vent frustrations, share their opinions on volatile subjects, or just go to share their life with the people who follow them, the amount of data created on these platforms has increased, leading to the use of a technique known as Sentiment Analysis. Twitter will be used as the platform for the analysis, this is because every second, on average, around 6,000 tweets are tweeted on Twitter. (internetlivestats.com, n.d.)

The purpose of this project is to utilise what is available and use that to create an effective real-time Sentiment Analysis model through Natural Language Processing, that will take a post and classify it as positive, negative or neutral in sentiment. A computational framework is ideal here, as this will enable opinion mining and sentiment analysis which can adapt to the domain. (Technopedia, n.d.)

This project aims to utilise Machine Learning, as the fundamental advantages of this, offer different foundations for learning. Utilisation of Machine Learning will allow us to analyse data more efficiently and to a much higher degree of accuracy. The project will make use of Tensorflow and Keras for machine learning, and Tweepy to extract tweets via the Twitter API.

The Primary Objective here, is to classify textual data through modelling, and thus obtain a classification model, that will achieve a sentiment accuracy close to around 80% or more. Retrieval of data and model investigation are the key Secondary Objectives of the project.

Data will be taken from the Twitter API and will be cleaned of any URLs before being used in the analysis. The data will be scraped with a topic in mind to analyse, this way there will be a wide range of opinions and overall sentiment which can be measured and used to train the model effectively.

Model investigation involves the browsing the algorithms available and looking at which would be the most suitable is important to the goals of the project, achieving a high accuracy. Deciding on the best algorithm to use is a big part of creating the final model, use of a few different algorithms to find which can give the highest accuracy is important.

1.2. Scope

The project will include analysis of data and a machine learning model that aims to meet the objectives. The model will be trained on data that is found online with the Twitter API as the source, and perhaps another social media site where the data links to the Twitter data and helps the model come to a more accurate solution. The aims of the model will be to learn from the data and analyse it so that this data can be classified with a high level of accuracy.

The project will not use data that doesn't relate to the data found through the Twitter API and the model will achieve static sentiment analysis from past data. If time allows, the model will also be able to classify data in real time and give an accurate result.

1.3. Deliverables

The project will deliver a working Machine Learning model that is able to classify the data into clusters which can then be used for visualisation. When the model successfully achieves sentiment analysis accuracy, the results can then be visualised, and findings can be shown and delivered in the end report. This will allow completion of a report, that showcases how the model was trained to come to this conclusion by itself, discusses findings, the iteration stages of the model along with visualisation of key findings.

1.4. Constraints

One of the biggest constraints to consider, is Time. A good project plan accounts for the constraint of time, but its not set in stone, and there is also a deadline that must be met, which means time plays an important role in risk analysis, project planning and on the project.

The data for this project will be scraped directly from the Twitter API using a personal Developer account. It is of great importance that the data taken doesn't identify an individual, as this would violate GDPR and thus needs to contain no personal data that can be used to identify any individual. Making sure that the data that is being taken doesn't identify an individual user, will make sure that adhering to GDPR, has no real impact on the project. (EU-Commission, 2018)

It is also important to consider the factor of Tools, or the Hardware that may be needed to compute a large dataset in a considerable timeframe. Larger datasets can take considerably more computing time to achieve a result than that of a smaller, more compact dataset with less missing values and more complete data.

Something that is important to consider, is model inaccuracy, which can stem from computer programs simply being unable to decipher language concepts of Irony, or Sarcasm. The model may misclassify statements or text found in the dataset which could lead to inaccuracy towards the result. It is important to make sure the results are thoroughly checked, and to make sure any inaccuracies are caught at the first instance and outliers are dealt with correctly wherever they may lie. (Guess, 2011)

1.5. Assumptions

The most important assumption is that there is not enough time to complete the project, or to achieve the secondary objective, and this is one of the reasons that an effective project plan is important. Working to a deadline means that this assumption is justified, as time will always be a factor and will mean that the plan of the project should always be considered at every step, making sure that it is always considered if a step in the project can be done quicker than is initially anticipated.

The assumption is that there will be no hardware failures in the personal workspace. This has a high likelihood of being true but making sure that there are plans in place in case there are failures is important to consider. This is something that is outlined in the risk analysis, and steps to acquire new hardware to work on should there be any issues will be discussed and agreed with my supervisor if the need arises.

2. Project rationale and operation

2.1. Project benefits

Sentiment Analysis is a great way for a business to find out what people think and how to decipher good and bad opinions of them. A successful project here would develop a model that can classify good and bad comments and visualise these effectively for a business. A high volume of comments towards a topic or brand is only good if most of these encompass positive sentiment.

One of the benefits of this project would be perception. Businesses could use the model to look at comments and see if they are positive or negative and then use them in their PR, as they can ensure that the messages they are trying to send, comes across in the right way and will reach the right audience. (Marta, 2019)

Positive sentiment is also important for people with standing, be that a person in music or a politician. These people who are at the top of their relative fields will care about their image, and this model will help them to understand what people think of them and give them opportunities to do things that help to improve their overall image in the real world and on social media platforms that they use.

2.2. Project operation

The project will operate with an Agile like Methodology, that will apply Agile like mechanisms into Incremental prototyping. The incremental model is a good choice here, as the elements of the model could be created as an Increment, and the after analysis and evaluation of the model, if the feeling is that the model could be improved, these steps could be done again. (Guru99, 2019)

It is better that the project doesn't follow a Software Engineering methodology such as the Waterfall Model, as there will be times where there is a need to look at the data again, or there may be instances where it is necessary to refer to the literature review, and this means taking steps backwards to move forwards. Therefore, a more Agile like approach to this project will be beneficial for a smooth conclusion.

The project will be running in 6 stages;

- Literature Review
- Data Analysis
- Algorithm Development
- Result Analysis
- Review
- Final Report

Using this stages method, an Agile approach is key, so that the project can go back and forth through the stages where necessary, if time permits this to happen it is important to make sure that back and forth navigation through these stages has a minimal impact on the time that is allotted for certain tasks, to keep the project on schedule always.

The best way to measure the success of this methodology choice, is time management, making sure that any changes to current project stage is documented and doesn't impact the critical path in anyway. Causing issues towards the critical path of the project always needs to be considered , and this is the key element that is kept under consideration throughout.

The benefit to this style of development strategy, is that it allows the project to be able to generate results quickly, whereas keeping with the flow and keeping to the deadlines set out. This is also an advantage as errors can be resolved quickly and then that aspect of the modelling can be redone without many problems. One of the downsides of this method though, is that iteration phases are rigid, and they do not overlap. This means that if the review stage is underway, its not feasible to go back to the initial modelling until the evaluation is done beforehand.

2.3. Options

There are many options when it comes to the design of the project. The main tool that will be used for this project is the Python programming language. The reason for this, is due to all the available tools that can be found within the Python framework, such as Tensorflow, Keras and Jupyter.

One of the tools that will be great for this project is the Jupyter Python IDE, most notable Jupyter Notebook. Jupyter itself, is an open-source software that allows you create dynamic Python code that can be spaced out into individual code blocks for easy integration and easy to view results for snippets of code. It works exceptionally well with the best libraries for Python used for data science.

In Data Science there is a great opportunity to choose from and use many great libraries that help to make the work of analysis much easier. First of all is Pandas, which is a great tool used for Data Munging and Preparation, it is a tool that helps to clean the data before analysis, eliminating missing values and helping to make sure that our analysis will be accurate and will give us the results that we expect.

For the machine learning aspect of this project, the best options are Keras and Tensorflow. Tensorflow is a fantastic library for Python created by Google for Machine Learning, whilst Keras is a library that runs on top of Tensorflow as it is more streamlined and has ease of use. Tensorflow is the best choice for the project, with its flexible and comprehensive ecosystem of tools, libraries and resources available, it will allow the project to run smoothly and be able to deploy a suitable machine learning model.

For the visualisation aspect, the project will take advantage of the Seaborn library, which is a library that offers ease of use and runs on top of Matplotlib. This is similar in functionality to that of Keras, it is a library that helps to create the implement the same great visualisation techniques, whilst also being ease of use. The last library in use here is Tweepy, which is the library that allows access to the Twitter API for extraction of tweets of yourself or that contain a word or timeframe of your choosing. This is important to apply machine learning to topical tweets towards sentiment discovery.

2.4. Risk analysis

Risk	Likelihood	Severity	Impact	Mitigation	Adjusted Likelihood	Adjusted Severity	Adjusted Impact
Data Difficulty	Low(2)	High(4)	8	Ensure data is backed up from the start and regular backups are kept reducing possibility of data loss	Low(1)	N/A	4
Inability to get training data	Low(1)	Very High(5)	5	Seek alternatives in case scraping method yields bad results, or developer access is denied or revoked	N/A	Medium(3)	3
Hardware Failure	Low(2)	High(4)	8	Ensure replacement hardware is available at short notice and the notebook file is saved to GitHub regularly	Low(1)	Medium(3)	3
Personal Illness	Low(1)	Medium(3)	3	Time for each section of project plan is slightly over the amount of time required to make up for any sudden illnesses	N/A	Low(2)	2
Model Inaccuracy Issues	Medium(3)	Medium(3)	9	Allow additional time for the model to be trained to account for any inaccuracy concerns	N/A	Low(2)	6
Assignment Deadlines affect time allocation	High(4)	Very High(5)	20	Ensure efficient planning of modules outside of the project to account for expected workload	Low(2)	High(4)	8
Loss of Twitter Developer Account access	Low(1)	High(4)	4	Save all scraped data into an initial CSV file that can always be accessed afterwards to mitigate risk of account loss	N/A	Low(2)	2

2.5. Resources required

The project is not expected to need any resources that are not already readily available for a Computer Science student account. Any software that would be required can be provided through this account, and most software required for this project is available open-source and is free to use.

The equipment needed for this project, is a computational device, such as a PC, that can run the calculations in a timeframe that is suitable and works within time constraints. In the case of equipment, I have the availability of a personal computer that is fast enough to deal with these computational tasks. There is also availability of PCs to use within the university, should any problems arise with my own hardware.

3. Project methodology and outcomes

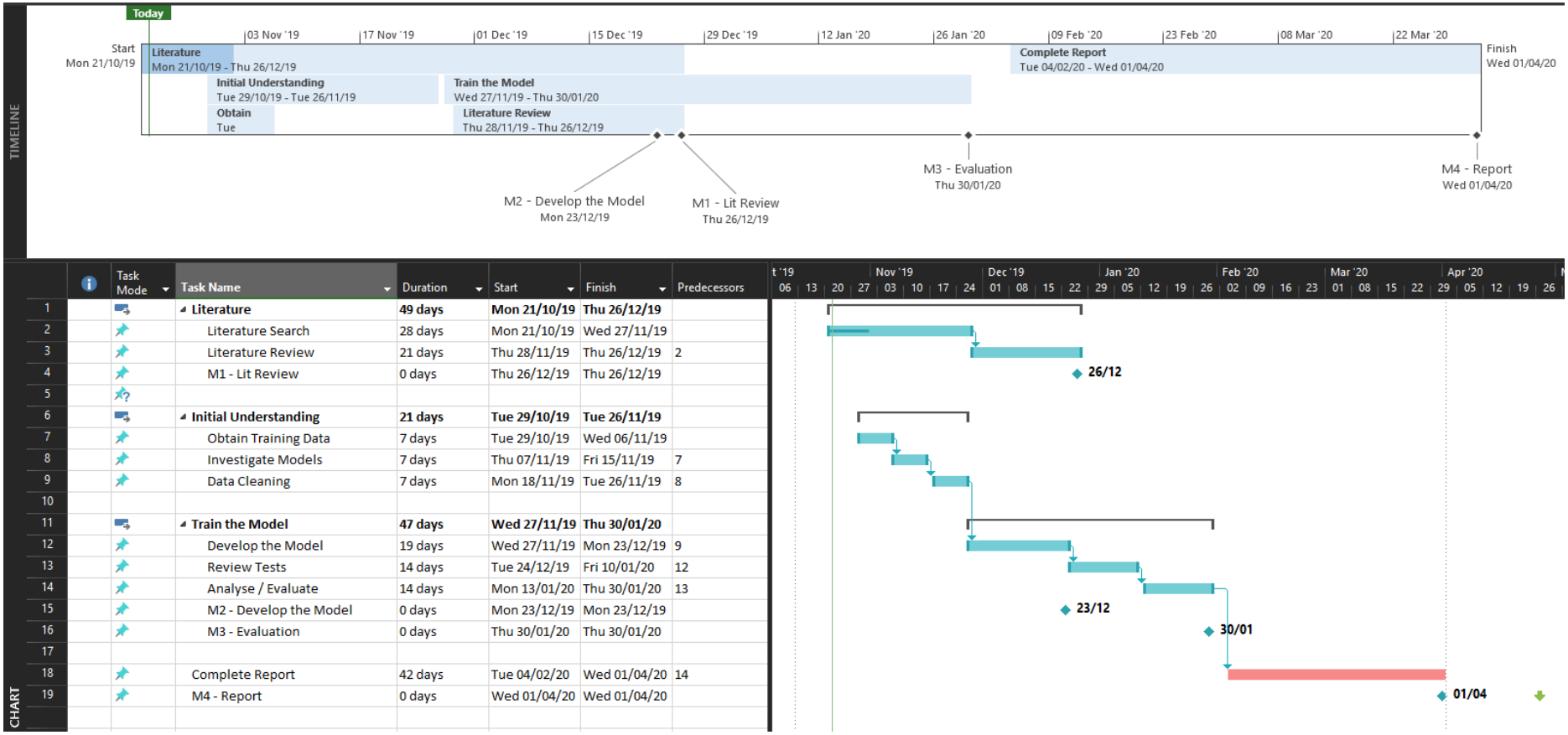
3.1. Initial project plan

3.1.1. Tasks and milestones

Task	Content	Time to Complete
Literature Search	Search for relevant books, articles and websites that have relevancy for the subject. Important information will be extracted.	28 days
Literature Review Milestone	This will involve reviewing of all-important data found and compiling it into an easy to use list relevant for the final report.	21 days
Obtain the Training Data	Training data needs to be found to begin the training of the model; this data also needs to be checked for suitability.	7 days
Investigate Existing Models	Investigate strategies for how models are developed to reach a suitable conclusion as to which model will be best.	7 days
Clean the Data	Clean the data of things that may ruin the accuracy of the implementation.	7 days
Development of the Model Milestone	Develop the model using the chosen methodology and aim to reach a classification accuracy before reviewing any tests.	19 days
Review Model Tests	Review the model tests to make sure they are achieving the results as expected. Make changes if there needs to be any changes.	14 days
Analyse & Evaluate Results Milestone	Analyse and Evaluate results, check classification accuracy, see if this can be improved further and evaluate as necessary.	14 days
Complete Report Milestone	Completion of the report using Visualisation techniques and presentation of findings from test results and model completion.	42 days

3.1.2. Schedule Gantt Chart

The Gantt Chart is shown below, with a task list attached to show dates assigned towards tasks. The section of the chart shown in red is the Critical Path, which indicates which part of the plan will cause a delay in the completion of the project. The green arrow shows the expected assignment deadline, and this gives a visual representation of the distance between expected end-date and assignment due date.



3.2. Project control

The project will run day-to-day, with progress being measured through regular checks on the project plan, making sure the plan is kept up to date on progression of each section and that dates are adhered to. If any part of the project looks like it may run late, changes to the plan need to be made and this needs to be accounted for in the weekly log of events.

A weekly log will be made which will briefly discuss the current state of the project, where the project is now, what progress has been made towards the existing plan, and anything that has come up that wasn't foreseen at the start of planning.

The project will be measured for success through these weekly logs and through the progress of model accuracy and progress towards the end goal of achieving the accuracy rate of 80% or more.

4. Appendix a: References

EU-Commission, 2018. *European Commission.* [Online]
Available at: https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules/eu-data-protection-rules_en
[Accessed 19 10 2019].

Guess, A. R., 2011. *Dataversity.* [Online]
Available at: https://www.dataversity.net/the-possibilities-and-limitations-of-sentiment-analysis/
[Accessed 12 10 2019].

Guru99, 2019. *Guru99.* [Online]
Available at: https://www.guru99.com/what-is-incremental-model-in-sdlc-advantages-disadvantages.html
[Accessed 18 10 2019].

internetlivestats.com, n.d. *Internet Live Stats.* [Online]
Available at: https://www.internetlivestats.com/twitter-statistics/
[Accessed 11 10 2019].

Marta, 2019. *Brand24.* [Online]
Available at: https://brand24.com/blog/the-benefits-of-sentiment-analysis/
[Accessed 12 10 2019].

Technopedia, n.d. *technopedia.com.* [Online]
Available at: https://www.techopedia.com/definition/29695/sentiment-analysis
[Accessed 11 10 2019].