

## Project Diary

### TR1 Week 2 (17/09/20)

#### Meeting notes

This first meeting was a group meeting, so only covered the more general admin side of the project - we went over the basic structure it could take, some of the ways of keeping on track and managing time, and loosley covered the development cycle of iteration and testing.

### Week 3 (23/09/20)

#### Progress since last meeting

I had created a rough draft of the IPO, and made a kanban board on Github to sketch out the project. I had also done some initial reading around fake news and the importance of headline.

#### Meeting notes

This was the first one-on-one meeting initially to cover the basics of the project. We talked about some next steps, including making a Gantt chart, getting the dissertation template set up with relevant headings and maybe looking in to either sourcing a dataset or creating my own.

I also got more of an idea over how to go about implementing the algorithm, and initially I'm feeling more drawn to a natural language processing approach as opposed to a machine learning one - it won't necessarily require a labelled dataset, and could be quite effective at detecting incongruence. Of course, I'll need to do more reading around this to settle on a specific approach.

### Week 4 (30/09/20)

#### Progress since last meeting

I created scripts to scrape the BBC and The Guardian for the top daily news stories, as well as one that collected the BBC's 'On This Day' archive, roughly 2000 articles from 1950-2005. I've started to write about the process of data collection in the dissertation. I also created a Gantt chart for the project, detailing how I plan to spend my time and the various deadlines I'll need to meet.

#### Meeting notes

We covered quite a bit in this meeting. I went over the data collection work I'd done, and Simon mentioned that it would be nice to have a good example of an article who's headline doesn't match the body at all, and to go through it manually to see if the human process of spotting the incongruity could be implemented in an algorithm.

Simon suggested a few NLP libraries I could look into - NLTK and SpaCy being prime candidates for this work. I also talked about my reluctance around reading articles and starting the literature review, and Simon gave some advice for reading and avoiding procrastination

Possible future work was also discussed - the real world implications of misinformation, creating a tool to help individuals verify the articles they're reading aren't trying to mislead them, and also looking at how language use predisposes reactions, for instance how car collisions with bicycles are predominantly described as 'accidents' in the media.

## Week 5 (7/10/20)

### Progress since last meeting

I did some very rough initial research, collecting a few articles and collating them in the dissertation document. I also added some headings and cleaned up the template a bit. I found The Independent's full archive, and set up a Python script to slowly scrape through an estimated 300,000 articles.

### Meeting notes

I talked about the possibility of creating a small one-page site to allow anonymous volunteers to label the dataset, although I was conscious of the time. Simon commented that while a labelled dataset is always useful, if I'm going with NLP then it's not a requirement. However, as we're so early on in the project, there is time to go down a few routes like this, so I will probably make a quick prototype of the site.

We also talked about the rough draft I had created, and the need for an introduction and good examples of the problem I'm trying to tackle.

We briefly covered summarisation as a technique to help analyse big chunks of text, but I'll have to do some tests to see whether the meaning of the article is lost once condensed.

## Week 6 (14/10/20)

### Progress since last meeting

I've collected 45k articles so far, so am starting to have a healthy dataset to work on. I've also created a very basic html skeleton for a data labelling website, and I've added some examples of fake news, sensationalism and clickbait to the lit review.

### Meeting notes

We covered possible future applications of the work, such as detecting personal information, monitoring hate speech and generating a measure for the trustworthiness of a website.

We discussed a lot about the data labelling side of things, too. Simon suggested that another avenue could be relabelling an existing dataset, although as I'm in the process of building one up at the moment it makes more sense to label the new one. Deciding a good subset of the data to label posed a problem - it'd be good to get more than one rating for each, so an average could be taken. If the subset's too large, then not enough ratings may come in, if it's too small then there won't be enough data to provide a meaningful evaluation. During the meeting, I had the idea to also time how long someone takes to read an article, as there could be a wealth of trends to be discovered, but at the same time I don't want to collect more information than I need. I also asked Simon for a VM to host the labelling site on.

In the coming week, I'll aim to get the labelling site in a publishable state, as well as working more on the lit review.

## **Week 7 (21/10/20)**

### **Progress since last meeting**

I've made progress on the lit review, adding in sections around existing approaches and fleshing out the data collection section. I also completed the backend of the labelling site, so it's almost ready to go live and start labelling the dataset.

### **Meeting notes**

We briefly covered the different NLP approaches, and the limits of each. Simon gave me some pointers on how to get started with the more domain-specific part of the research, as well as mentioning how argumentation research into determining relevance could be appropriate for this project.

We also covered the process of gaining ethical approval for the data labelling site.

For next week, I'll try to work more on the NLP side of the lit review, as well as filling out the required paperwork for ethical approval.

## **Week 8 (28/10/20)**

### **Progress since last meeting**

I completed the ethics documentation and sent it to Simon for approval. I've also got the labelling site into a publishable position, and have created a 300-article dataset ready to be labelled. I also updated the lit review with a section on NLP and tidied up some previous sections.

### **Meeting notes**

The server Simon set up for me has issues, so I will use my own for the labelling site. We also talked about the cleanliness of the dataset - from collating the 300 articles for labelling I found about 10% of them had been incorrectly scraped and weren't comprehensible. We covered maybe using a spell checker to weed out the dirty articles from the main dataset, or just leaving them in and seeing how the algorithm handles them.

We also talked about the interim report, what should be included in it, and the rough structure of the dissertation.

For next week, I'll set up the labelling site on my own server, add an introduction to the report, and do some more work on the NLP section to cover structured and ML approaches.

## **Week 9 (4/11/20)**

### **Progress since last meeting**

I've made the labelling site live and shared it with a range of groups. I've managed to get 150 individual article ratings so far. I've also tidied up the first part of the lit review and started working on an introduction.

### **Meeting notes**

We covered the 'bad' data generated by the labelling site (very poor scores that took 2-3 seconds to rate) and decided to leave them in the dataset to allow for greater analysis. We also talked about the

challenges I'm facing around getting to grips with NLP, and the breadth and depth of the subject, as well as the different approaches I could take to building the classifier. For next week, I'll do some more work on the lit review and start to get comfortable with an NLP library so I can begin to create the classifier.

## **Week 10 (11/11/20)**

### **Progress since last meeting**

This week has been less productive than previous weeks, but I was able to tidy up the dissertation in preparation for the interim report. I also started to look at sentiment analysis and how it could be used to tackle the problem.

### **Meeting notes**

Due to time constraints this was a shorter meeting than usual. We covered my concerns about not having the skillset to do the task - NLP isn't something I've come across before in my studies. Simon suggested a range of NLP technologies to look into, such as word2vec, bag of words, and one hot encoding. For next week, I'll read about these other NLP approaches, and start to think about creating the classifier.

## **TR2 Week 2 (26/01/21)**

### **Progress since last meeting**

Over the Christmas break I worked on tidying up the current state of the dissertation, adding some basic framework to the sections and formatting the experiments. I also researched and wrote up sections on lexical similarity and tf-idf. I was able to do an experiment with tf-idf, but the results showed it wasn't suited to detecting congruence.

### **Meeting notes**

We talked about the work I had done over the break, and I raised the point that all the experiments I'd done so far had failed, in the sense that none of the approaches would help with the overall project's aim. Simon mentioned this wasn't a fatal problem in the project, and clarified that the aim is not to answer the question but to attempt to answer it. We talked about next steps, such as looking at summarisation, and doing analysis on the dataset I had collected so far.

I also made the decision to stop work on the labelled dataset, even though not enough responses have been collected - it is now unlikely the labelled dataset would be used as a classifier may not be built, and there are more important things in the project that require my time.