Project Diary

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 initally 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.