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.

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 concious 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 analyise 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 exisitng 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.