11/10/2019 Meeting 1

1. **What did you envision for the project when you created it?**
   - **Potential project themes?**

   - **Potential direction of project?**
     Cutting edge NLP, the accuracy of the finished tool is not so important, try and compare techniques with full knowledge

   - **Potential research questions?**
     Very different from the project title
     i.e. If the project title is "machine learning for fake news detection", the research question should address the imbalance problem (about your contribution in solving the problem)
     More specific than the project title, highlighting the challenges
     Not set in stone (the question can develop throughout the project

     Example 1: Investigation of contextual embedding with this problem and how to visualise it
     Example 2: How can you detect the words that correlate more to certain labels

2. **Do you anticipate that we should look to be more specific with the focus of the project in terms of the types of anomalies we should detect? Is the project title as stated in the project list?**
   - Yes, should be more specific as project title is too broad. For example, "fake news detection"
   - The project title should be changed to something more specific

3. **Could we discuss deliverables for the project in terms of basic, intermediate and advanced and out of scope? As an example, where would deep learning approaches fall?**

   Basic:
   - Find a data set and apply traditional NLP techniques like frequencies and n-grams for Latent Dirichlet Allocation (lda), Svm, and produce some accuracies
   - Cleaning and segmenting data

   Intermediate:
   - Data that we create embeddings with word2vec
   - topic modelling and visualising the embedding
   - apply deep learning classifiers

Advanced:

Including contextual embeddings and deep models
Not just identifying if news is fake or not fake, it should also be able to identify specifically which part of the news is fake

4. **Is the finished product a "commercial" tool? Should it incorporate a GUI?**
   Command line code - GUI not necessary
   Basic interface at the end as an advanced deliverable (i.e. if you have time)

5. **Should the tool work for languages other than English?**
   Just for English (there is lots of data available for English) English models are very accurate and many other languages have difficult patterns
   Also, uni would need to bring in an external marker to test the accuracies in other languages

6. **How do you feel about the use of machine learning libraries such as TensorFlow and PyTorch, is this oversimplifying development?**
   Yes we can use these libraries (just using existing libraries - it is better to reuse than reinvent the wheel)
   SpaCY, Scikit learn, Keras (soft deep learning library) - would recommend Keras

7. **How much of the tools development *process* is included in the report?**

8. **Are labelled or unlabelled data sets better?**
   labelled data sets are better (labels could be offensive or non-offensive)

9. **Should the final tool include more than one anomaly detection technique, or just the best identified approach based on research?**
   During research, mixing techniques and trying different things
   In the final paper, they are separated

Previous students
Jack Layland
Thomas Winterbottom
Dean Slack
Thomas Hudson

Noura's GitHub username: NouraMoubayed

Research Questions

1. **Do you have any pointers for areas of research?**

General Questions

1. **Can we cite sources such as articles on websites such as medium.com? Should they instead by academic papers?**
   Academic papers only

2. **What do you envision should be complete by our next meeting**
   a. Produce a list of datasets e.g. *google toolbox* or potentially crawl data from twitter if you are looking at a specific application
   b. Rank them for your favorites (summarise with x and y) give some stats about the data and summarise
   c. Read up on SpaCY and sci-kit (SpaCY is most important)
   d. Make a GitHub repository with folders for each deliverable (e.g. lit survey, design report…)
   e. make a powerpoint with a slide for each week - make sure to present the information nicely (visually appealing), maybe with pie charts about dataset for first slide
   f. Include a list of interesting papers
   g. Start lit review

3. **Roughly what is the percentage breakdown of how much of the project should be about the research and how much should be about the tool?**

My Ideas about how the project could work:
Structure 1:
Research an approach
Implement an approach
Experiment and produce results
Analyse results
Repeat and compare

Structure 2:
Research lots of approaches, comparing other people's results
Make an objective judgement of which will likely be best
Develop the tool with the chosen approach