# Initial Project Plan

## Hyperpartisan News Detection

CM3203 - One Semester Individual Project

40 Credits

Student: Patrick Noyau

Supervisor: Luis Espinosa-Anke

# Contents

# Project Description

**The main aim of this project is to develop a system to predict if a given news article is hyperpartisan (especially left-wing or especially right-wing).**

Google searches for the term "fake news" increased by over 300% in November 2016, during the time which the United States Presidential elections were taking place (Google Trends). Popularised by then presidential candidate Donald Trump, the term is now almost ubiquitous, and describes social media posts or news articles designed to spread disinformation (Wendling 2018). The ease at which these falsehoods can spread can have a substantial impact in democracies, especially around election periods. (Bakir and McStay 2018) suggests that 'fake news' produces uninformed citizens, that stay uninformed due to the echo-chamber (only encountering opinions that coincide with your own) effect of social media. Citizens may also become "emotionally antagonised or outraged" as much of the falsehoods produced are designed to be provocative in order to be effective.

Research by (Silverman et al. 2016) categorised articles published to Facebook from selected mainstream, left-wing and right-wing publishers, into the categories; mostly true, mostly false or a mix or both.

| **Publisher type (articles analysed)** | **Mostly True** | **Mostly False** | **Mix** | **Combined Mix and Mostly False** |
|---|---|---|---|---|
| *Mainstream (826)* | 97.6% | 0% | 0.969% | 0.969% |
| *Left-wing (256)* | 71.1% | 5.86% | 19.9% | 25.8% |
| *Right-wing (545)* | 50.6% | 13.2% | 8.07% | 21.3% |

Fig 1. Summary of categorisation. Excludes articles with "no factual claim or content" (Silverman et al. 2016)

The results showed that, although the majority of articles from all publisher types were categorised mostly true, articles from mainstream publishers were far more reliable (97%+) at being mostly true. At least a fifth of articles from right-wing publishers contained some falsehoods, as with a quarter of articles from left-wing publishers. Comparatively, less than 1% of articles from mainstream publishers contained any falsehoods and none were found to be mostly false.

There are three different possible ways of identifying fake news; knowledge based detection, context based or style based detection (Potthast et al. 2017). Knowledge based detection essentially compares statements in the article to factual knowledge, checking for inconsistencies. Context based detection seeks to analyse how fake news spreads on social networks in order to create algorithms to counter the spread. Style based detection works by analysing articles known to contain false information to see if they have any overlap in writing style. New articles can then be compared to known fake news writing styles to see if they may also contain false information.

In the paper *A Stylometric Inquiry into Hyperpartisan and Fake News* (Potthast et al. 2017), a series of experiments are performed using style based detection to identify if there are common style features between fake and real news, and also between left-wing and right-wing articles. Using unmasking, a process where the features that best distinguish between two articles are iteratively removed, to see the speed at which cross-validation accuracy decelerates (Koppel et al. 2007). When unmasking was applied to left-wing and right-wing articles cross-validation accuracy decreased rapidly, meaning that the style of the two types of articles are very similar.

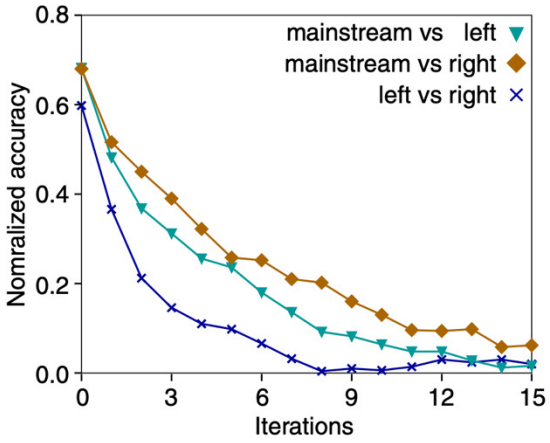


Fig. 2. Unmasking applied to pairs of political orientations. The quicker a curve decreases, the more similar the respective styles are. (Potthast et al. 2017)

Given the similarity in style between left-wing and right-wing articles, we can combine the two types into the catchall term, 'hyperpartisan'. Hyperpartisan articles were shown to contain a far higher rate of falsehoods compared to mainstream ones, therefore if a classifier can be made to detect if an article is hyperpartisan, it would be a useful tool in combatting fake news. As part of SemEval-2019, an International Workshop on Semantic Evaluation, a contest was held where teams tried to produce the most effective classifier at detecting

hyperpartisan articles (Kiesel et al. 2019). A dataset of manually labelled articles were given to participants to train a their classifiers.

“ Hillary Clinton’s niece is turning her back on her “ selfish ” aunt — to endorse Donald Trump for president .“ I support Donald Trump 100 percent , ” Macy Smit , the daughter of Bill Clinton’s half-brother Roger Clinton , declared to RadarOnline .“ I have been a Democrat my entire life , but Trump is what we need right now — somebody who is going to stand up for us . I think at this point Hillary just wants it for the history books — to be the first woman president for selfish reasons , ” the Tampa hairstylist continued . Smit , 25 , admitted she’s never met Hillary Clinton , and said that side of the family doesn’t think much of her .“ Something tells me the Clinton side looks at me and my mother as not good enough , but we’re hard-working , ” Smit told the gossip site . Roger Clinton was a deadbeat dad who dumped Smit ’s mom , Martha Spivey , shortly after she became pregnant with his child , according to RadarOnline . “ The Clintons are all talk ! ” Spivey , 50 , told the site .“ Hillary says she’s all about family , but she’s got a niece she’s never met and never acknowledged .The Clintons have never helped us out. ” Hillary’s lost the trust of even more voters with her inconsistencies : ”

Fig. 3 - A hyperpartisan article from the dataset

“ President Donald Trump announced Wednesday that his administration “ will be taking strong action today ” to address what he complained are especially weak border security protections that must be addressed by Congress and the construction of a border wall .“ Our Border Laws are very weak while those of Mexico & amp ; Canada are very strong . Congress must change these Obama era , and other , laws NOW ! ” Trump wrote on Twitter . “ The Democrats stand in our way - they want people to pour into our country unchecked ... .CRIME ! We will be taking strong action today. ” Trump’s online post , which did not come with further details as to what steps his administration would take , came one day after he announced that he would deploy U.S. troops to guard the southern border with Mexico until his promised border wall is completed . White House press secretary Sarah Huckabee Sanders said later Tuesday that Trump’s plan would mobilize the National Guard , which his two immediate predecessors had also done along the U.S.-Mexico border , and also press Congress to take steps to close “ loopholes ” in the nation’s immigration laws . Sanders did not say when or how many troops Trump would mobilize or where they would be deployed . The president has been especially preoccupied with border security this week , seemingly agitated by a group of migrants , mostly from Honduras , that was making its way north through Mexico , past immigration checkpoints and military bases /, en route to the U.S. Trump said Tuesday that that group had been broken up by Mexican authorities , who he said did so only after he demanded they act . More broadly , Trump has struggled to get traction in Congress for his border security priorities , namely that the border wall he promised during the 2016 campaign would be paid for by Mexico . Democrats have thus far been largely unwilling to go along with plans for the president’s wall . Trump has complained loudly about what he has called their obstructionism and has urged the Senate GOP leadership to do away with the chamber’s legislative filibuster , something Senate Majority Leader Mitch McConnell has vowed never to do . ”

Fig. 4 - A non-hyperpartisan article from the dataset

# Aims and Objectives

The aim of this project is to produce an effective categorisation system that takes text from a news article as an input and outputs a binary number, based on if the news article is categorised as hyperpartisan or not. Through creating a classification system for hyperpartisan news, I should also be able to identify key features that hyperpartisan news articles share amongst them. As discussed previously, hyperpartisan articles are far more likely to contain false information, so the ability to identify where articles are hyperpartisan will provide a useful tool in combatting the spread of fake news.

Over the course of this project, these are the objectives I aim to achieve:

- To be able to understand the different machine learning methods for text categorisation
- To identify which machine learning method(s) will work best for hyperpartisan news detection
    - Use findings from contest entries in (Kiesel et al. 2019) as a guide for what can be effective
- To implement the machine learning system in Python
    - To be able to understand and use different machine learning libraries in Python, specifically *scikit-learn*
- To create a classification system that can take a news article as an input and output if the article is predicted to be hyperpartisan or not
- To identify which features best discern hyperpartisan news articles from mainstream news articles
    - To produce a categorisation system that detects hyperpartisan news with the best possible accuracy, precision and recall

# Work Plan

There are several tasks that are required to complete this project, listed below. On the next page is a Gantt chart showing the expected completion dates. The project deadline is 7th May, and the aim of this time plan is to give a week at the end of the project for proof reading and final improvements.

Tasks

- Produce initial report describing the project and its aims.
- Begin to research machine learning algorithms and methods for text classification.
    - Using the research by (Kiesel et al. 2019), see how the teams completed their submissions and which machine learning techniques they used.
- Experiment with machine learning libraries in Python and try implementing different models
    - Follow tutorials for *scikit-learn* to strengthen knowledge of this library
- Create a classification system in python that takes a news article as input and categorizes it as hyperpartisan or not
    - Implement Python script to take input from news article and identify if it contains certain key features
    - Create / train a classifier to identify if an article is hyperpartisan from the features it contains
- Test the classification system and, as necessary, improve to produce the best possible accuracy, recall and precision.
    - Record performance data from experiments for discussion in final report
- Produce final system that is ready for demonstration at the Project Viva
- Write the final report, including the following sections:
    - Introduction
    - Background
    - Approach
    - Implementation
    - Results and Evaluation
    - Future Work
    - Conclusions
    - Reflection on Learning

## Gantt Chart

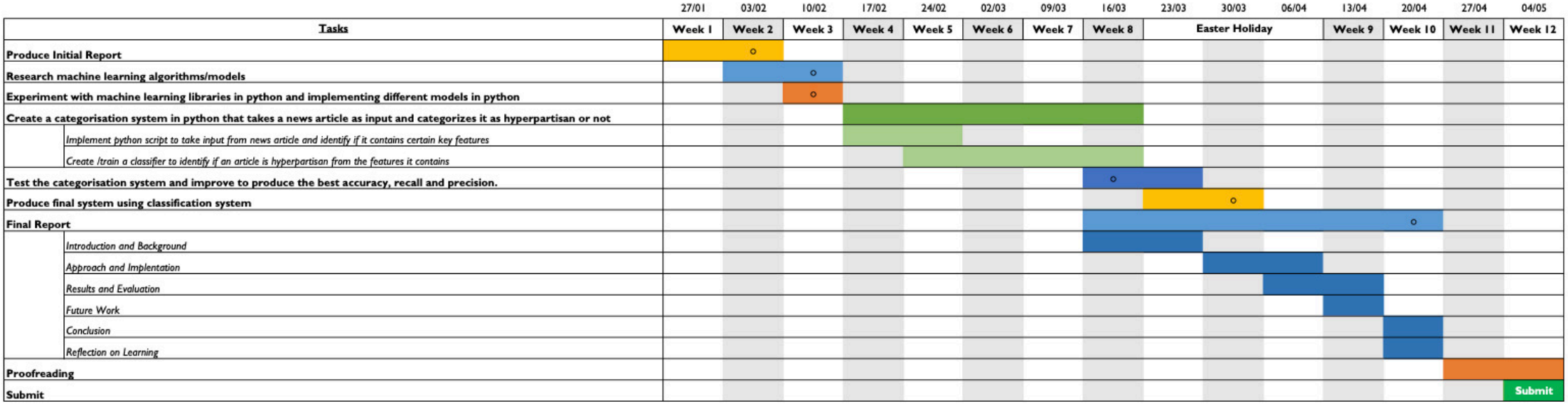


Fig.5 Gantt chart showing project timings. Also available separately.

Milestones

Below is a list of key milestones in the project, and the dates that I expect them to be met. Meeting these milestones on time will be a key indicator that the project is running to time.

| **Milestone** | **Date** |
|---|---|
| Submit Initial Report | Beginning Week 2 |
| Finish researching and experimenting with machine learning libraries | End Week 3 |
| ***Supervisor Review Meeting*** | Week 5 |
| Begin testing classification system and collecting experiment data | Week 8 |
| ***Supervisor Review Meeting*** | Week 8 |
| Have final system ready | Mid-Easter Holidays |
| ***Supervisor Review Meeting*** | Week 9 |
| Finish writing final report | End Week 10 |
| Submit Project | Week 12 |

# References

*Google Trends: "Fake News"*. https://trends.google.com/trends/explore?date=all&q=fake%20news: Google. Available at: [Accessed: 28/01/2020].

Bakir, V. and McStay, A. 2018. Fake News and The Economy of Emotions: Problems, causes, solutions. *Digital Journalism* 6(2), pp. 154-175. doi: 10.1080/21670811.2017.1345645

Kiesel, J. et al. eds. 2019. *SemEval-2019 Task 4: Hyperpartisan News Detection*. *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA, jun. Association for Computational Linguistics.

Koppel, M. et al. 2007. Measuring differentiability: Unmasking pseudonymous authors. *Journal Of Machine Learning Research* 8, pp. 1261-1276.

Potthast, M. et al. 2017. A Stylometric Inquiry into Hyperpartisan and Fake News.

Silverman, C. et al. 2016. *Hyperpartisan Facebook Pages Are Publishing False And Misleading Information At An Alarming Rate*. https://www.buzzfeednews.com/article/craigsilverman/partisan-fb-pages-analysis: Buzzfeed News. Available at: [Accessed 03/02/2020].

Trends, G. *Google Trends: "Fake News"*. https://trends.google.com/trends/explore?date=all&q=fake%20news: Google. Available at: [Accessed: 28/01/2020].

Wendling, M. 2018. *The (almost) complete history of 'fake news'*. https://www.bbc.co.uk/news/blogs-trending-42724320: BBC News. Available at: [Accessed 03/02/2020].