Cardiff University
Computer Science and Informatics

CM3203 - Final Year Project

Artificial Intelligence / Machine Learning for Understanding Misinformation in Social Media

Initial Plan

Author: Joe Harris

Supervisor: Dr Alun Preece

Contents

1 Project Description

With 67% of the UK's population consisting of active social media users [**1**], we have an unprecedented level of user-created data freely ready to analyse. Turning this big data into meaningful, comprehensive insights that tell us about the world at large is a hot topic in machine learning, with technological juggernauts such as Twitter now being able to advertise themselves as 'where people come to discover what's happening' due to the sophistication of their AI's insights [**2**].
This project looks at creating situational understanding from this stream of social media data, allowing us an insight into popular trends at given times, how these trends augment through time, and how this can inform us about the world in general.
A large part of this project will also pertain to the existence of misinformation within these social media streams, focussing on how this misinformation is conceived, the various techniques used to spread it through the social media populace, and what we can learn from all of this, in an effort to restrict its ubiquity and poignance as we move into a hopeful future of fact over fiction.
To discover this misinformation, I will be analysing how to detect a 'soft fact', which is a fact involving imperfect and ambiguous knowledge - a perfect vessel for misinformation.[**3**]
In this project I will be using a dataset of approximately 1.3 million tweets, collected during a period between April 2019 and June 2019, with the tweets relating to the EU election (achieved by searching based on specific terms). The dataset was obtained by the Cardiff University Crime and Security Research Institute, whom are hosting the project, and are providing myself with support and hot-desking facilities throughout the project.
This dataset has not before been analysed by the institute, and so the results of this project will allow an original insight into this period of time.

The goal of this project is to create a system that can process social media data, analyse it, and then classify it, identifying misinformation based upon the related topics, the accounts posting, and the popular themes at that time in the social media.

2 Ethics

The data being analysed and used throughout this project has been collected by Cardiff University Crime and Security Research Institute, under the appropriate ethics framework of the institutes' research. This data has been ethically approved, and I have been given permission to use it by the institute. This will be the only data I use through the project, as I will not be attempting to obtain any by myself, meaning there's no potential extra ethical breaches I must consider.

3 Project, Aims, and Objectives

3.1 Aims and Objectives

With the limited time of only 12 weeks in which to complete this project, I have created two sublists of aims and objectives: guaranteed aims, and desirable aims. The guaranteed aims are the fundamental, minimum items which must be achieved in order to fulfil the project in its most basic sense. The desirable aims are those that will allow the project to begin having a real world impact, and I have allotted time in my work plan to achieve these aims, but will approach these only once the minimum viable goals have been executed.

The ***core*** aims and objectives of this project are:

C1. Process tweet dataset, finding most common words and hashtags per epoch

C2. Create classifiers, and classify tweets with multi-class text classifier

C3. For each epoch, have a displayed understanding of the major topics

C4. Within the dataset, have flagging system that flags posts that may be misinformation

C5. Combine these functionality to a dashboard wherein this data can be obtained

The ***desirable*** aims and objectives that I will *attempt* to fulfil within the project are:

D1. Create a list of accounts linked to misinformation for each epoch, with some form of classification of the misinformation being spread by the account, allowing us an insight into the primary sources of misinformation

D2. Link these sources of misinformation to the methods in which they are distributing this false narrative (be it through retweeting, replying to large accounts with misinformation, spam tweeting etc.), to inform us as to the most common and effective misinformation spreading techniques

3.2 Milestones

Milestone ID	Milestone description	Method used	Objective being achieved
M1	Frequent words and phrases for each epoch functionality	Pandas (python data analysis library)	C1
M2	Multi-class text classifier functionality	Scikit-Learn	C2
M3	Top 10 topics for each epoch functionality	Latent Dirichlet Allocation	C3
M4	Misinformation flagging and identification functionality	Natural language processing, and looking at followers/following for account (and other techniques) [**4**]	C4
M5	Identified misinformation being spread, and how it's being spread	Naive Bayes Classifier	D1/D2
M6	Dashboard with all functionality	Dash and Plotty Python Libraries	C5

4 Work Plan

Below is my week-by-week work plan for the project. This is obviously a rough estimate as to how events will occur, and the timeframe required for each task, which is why the details are vague. It is likely that I will backtrack through the process of ultimately creating the fully functional dashboard, however this should not eat too much into the timings provided below, as each functionality roughly adds on to the previous week's functionality, in a waterfall model style.

Week 1: Create initial plan; draft project plan; create python script to find most common words, hashtags, and bigrams for each epoch (completing milestone M1)

Week 2 and 3: Research and create multi-class text classifier using neural network; test tweet data in model and refine as necessary;

Week 4: Supervisor meeting to deliver milestone M1, and review progress with members of the institute; modify neural network to begin using account data alongside tweet data in classification (i.e. giving more credence to political tweets if we can derive from the account details that it is a politics based account) (completing milestone M2)

Week 5: Create top 10 trends for each epoch based upon the classifications derived for each tweet in the neural network, and create graph showing fluctuations in these trends' popularity across full timeline (completing milestone M3)

Week 6: Supervisor meeting to deliver milestones M2 and M3, and review progress with members of the institute; research into identifying bot tweets from tweet dataset; test out approaches observing their effectiveness and efficiency (completing milestone M4)

Week 7: Investigate into identified bot accounts, finding all relevant information on them within each epoch, and finding the misinformation that they are spreading

Week 8: Research into how the identified bot accounts are actually spreading misinformation (i.e. the methods they are using), and how effective these methods are (in terms of favourites, retweets, replies) (completing milestone M5)

Week 9: Research how we can use collated insights to predict the future, through how insights in previous epochs potentially influence trends in future epochs

Week 10: Supervisor meeting to deliver milestones M4 and M5, and review progress with members of the institute; create the dashboard in which these insights, and bot accounts are identified, with all relevant graphs and charts shown on screen dependant upon input json file (completing milestone M6)

Week 11 and 12: Collate all recorded data and writing into the final document

5 References

[1]
Reference type: Online
URL: https://www.statista.com/statistics/507405/uk-active-social-media-and-mobile-social-media-users/
Author: S. O'Dea
Title: Total number and the share of population of active social media and mobile social media users in the United Kingdom in Publish date: January 2019
Accessed: 30th January 2020

[2]
Reference type: Online
URL: https://blog.twitter.com/en_gb/topics/marketing/2017/twitter-is-where-people-come-to-discover-whats-happening.html
Author: Gordon Macmillan
Title: Twitter is where people come to discover what's happening
Publish date: 11th June 2017
Accessed: 30th January 2019

[3]
Reference type: Paper
Title: Briefing Paper: Digital Influencing Engineering
Author: Universities' Police Science Institute
Publish date: 23rd April 2018

[4]
Reference type: Article
Title: Disinformation and digital influencing after terrorism: spoofing, truthing and social proofing
Author: Martin Innes, Diyana Dobreva, and Helen Innes
Publish date: 25th January 2019
DOI: 10.1080/21582041.2019.1569714