Lab Task - Early Idea Drafts

What initial ideas do you have?

art - image / artwork recognition (online automatic copyright check on social medias?)

a form of human behaviour

↓

fake reviews on amazon detection - tone, punctuation, overly product enforcement (one profile spamming one review?)

↓

work related issues - what issues do we encounter daily?
↳ no documents to track team updates - automate this?

content moderation - tone detection instead of relying on user reports or trigger words that flag the system

potentially even long-term
↳ online persona evolution analysis, somebody posting more and more aggressive political posts, likelyhood of terrorism?

themes so far: - micro patterns (tone, punctuation)
- larger behaviour traits (fake reviews, personality shifts)
- consequences - detecting harm, extremism

It's like building a small behavioural pattern detector!
↳ uses real reviews, tweets or posts
↳ clusters or labels these based on tone or intent
↳ tracks users over time
↳ flags when prediction detects something is off
↳ basically a moderation system

Week 2 - further exploration of this idea

Datasets found: US Election 2020 Tweets from 15.10.2020 - 8.11.2020
↳ located using #DONALDTRUMP
#JOE BIDEN

What am I hoping to find? 1. Multiple tweets from individuals to analyse their behaviour
2. Users that tweet daily
3. Potentially politicians
4. No specific behaviours, only changes
5. a pattern, cluster, an outlier or any form of before and after

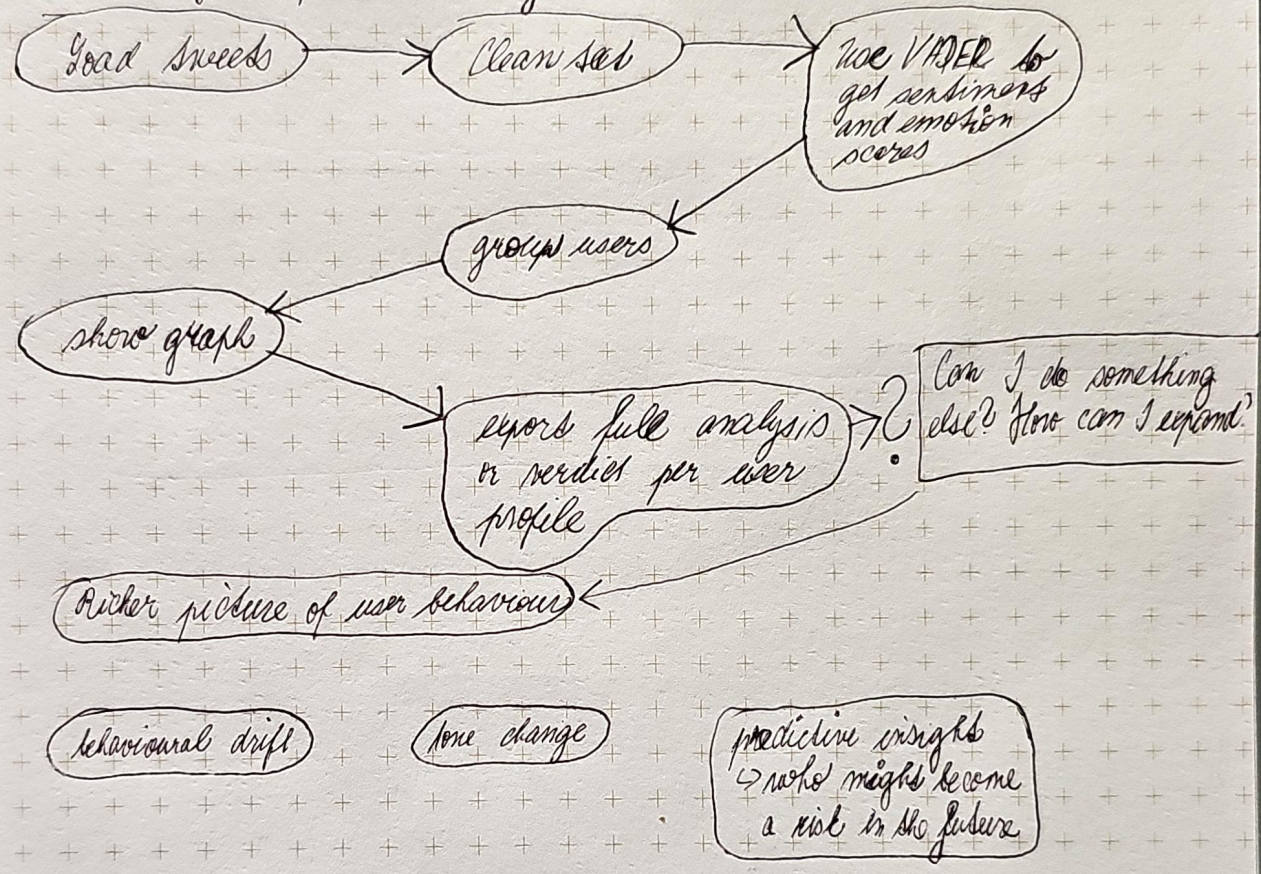
→ at first glance, this dataset is huge and seems to only have 10 tweets per person max → how do I filter it and utilise if only have such limited data?

Greg Andersen thesis notes: → extremism utilises propaganda (social media)
→ individuals opt to gain the sense of community support, info, resources
→ far-left uses emotions
→ far-right uses misinformation
↳ these could be used to determine ML parameters

Twitter sentiment analysis: Olga Kochyna et al.: → look user comments from before and after joining an extremist group
→ detected tone, behaviour changes
→ ML techniques used
↳ NLP to analyse text
↳ RANDOM FOREST to label or predict behaviours
↳ feature extraction

A review on sentiment analysis and emotion detection from texts (Nandwani, Verma)
↳ described how different researchers did sentiment analysis
↳ lexicon based methods (VADER or TextBlob)
ML models (logistic regression or support vector machines)
Deep learning
↳ VADER seems to be designed for social media tone detection

What does my workflow look as of now?



Project proposal notes

- ML system that analyses behaviour of a person
  ↳ revealing long-term behaviour patterns
- NOT only sentiment analysis! → it is about behavioural drifts
- hoping to find shifts in emotional intensity, tone and aggression, and potentially emerging risks and radicalisation

What I will use:
Pandas - to load, clean and manipulate my datasets

What exactly will I build?
- An app that loads a dataset (in this case a tweets dataset)
- processes on one or more users
- analyses sentiments, and potentially topics?
- detects behavioural shifts
- visualises shifts in simple graphs or outputs
- This can be later used in a risk assessment toolkit, or a tool used in academia in psych- mental health risk detection

What exactly am I doing with the data?
- read the datasets
- preprocess - clean the text - remove links, symbols, mention of the # used to collect this dataset
  - filter for users with 20+ more tweets?
  - sort tweets by time
- apply sentiment analysis - textblob?
  - Vader?
- group by user and time to see how behaviour evolves
- Visualise
- Potential other use of an extension - stop stressing, what you are doing is already ENOUGH

More academic papers

- Identifying and characterising behavioural classes of radicalisation within the QAnon conspiracy on Twitter notes
  - proposes a framework to detect and classify behavioural patterns used to radicalisation on Twitter
  - uses ~~tweets~~ user data (240 mil. tweets) to see how users engage with conspiracies and how they behave, and how effective moderation is → this is the basis of my project
  - What I found useful: - tracked behavioural signals
    1. Content Based Metrics
       ↳ How often users post about conspiracies
       ↳ Whether their profile info contains QAnon media
    2. Community Based Metrics
       ↳ How often user retweets a conspiracy
       ↳ Similarity in language style
  - Techniques used - K-means clustering + t-SNE (ML technique used to visualise high-dimensional data by reducing its dimensionality while preserving the relationships between data points)
    - defined 6 classes - amplifiers, self-declared supporters, hyper active promoters and more
    - used lexical similarity and posting patterns

- What are lexical scores doing?
  - measuring how similar users' language is to predetermined
    ↳ just need to define what language counts as radical

- How to source these definitions?
  - Use your own intuition and experience
  - Look for datasets focusing on radical/aggressive behaviour
    ↳ find bad words and build a framework using TF-IDF (term frequency inverse document frequency)
      ↳ measures how important a term is

- or use emotions/sentiments + heuristics ~~(again)~~ average using VADER
  ↳ track how this changes over time for each user
    - create a mark if sentiment drops, or anger related words appear, or if tweets become too emotional
- focus on signal based on a pattern

- Longitudinal sentiment analyses for Radicalisation research
  - investigates how sentiment changes over time in tweets related to radicalisation
    ↳ time window sentiments - tweets are grouped to per-week segments VADER is applied to each tweets → average sentiment is calculated
      - this shows sentiments trajectory over time
      ↳ I can use this! - Since I have about a month worth of tweets from the election period, I can group them by weeks.
  - clustering of sentiments - users are clustered by sentiments changes → some start positive and then drop, some the opposite, some stay stable, some are negative only
    ↳ maybe I can use K-clustering on users sentiment over time

- since this paper focuses on the Jan 6, it does a comparison I do not have acces to post election tweets
  ↳ maybe I can look up important pre-election dates from within my dataset

- Interaction-Based behavioural analysis of twitter social ~~media~~ network accounts
  - shows how to extract features like posting frequency and classify accounts into categories using KNN, SVM and ANN

- Moral foundations twitter corpus: a collection of 35k tweets annotated for moral sentiments - Hoover et al.
  - shows a large annotated dataset of tweets labeled with 10 moral categories
  - using TF-IDF and a classifier

Post feedback notes

- good idea, motivation and solid academic grounding, not generic
- VADER is not magic – it's only good for social tone, but what I need is something that can handle emojis, punctuation, caps lock, posting frequency
  - it doesn't "understand", it "reacts" only
- define labels better (no "negativity")
- score shifts are not emotional instability or radicalisation, unless I define these and expand on them
- focus on "user x shifts from X to Y" rather than labeling intent or ideology
- VADER only does "good=+1, bad=-1" based on words it knows
- maybe a better approach would be feature extraction + unsupervised model?
  - ↳ as mentioned – emojis, all caps, declaration points, swear words
- WHAT IF ⇒ using clustering I can group users based on features and then label them based on what I find?
  - ⇒ feature extraction → clustering → maybe anomaly detection? → then something that interprets the results? → visualisation
- unsupervised models like KMEANS or DBSCAN?
- I could implement some temporal behaviour – shifts per week/day?
- bag-of-words is a simple version of TF-IDF → only counts how often words appear
- add caps – learn

What should the new process look like?

1. Feature extraction:
   - Pandas for aggregation (take data and calculate statistics)
   - Regex for finding symbols, punctuation or swear words
   - Datetime for finding the posting frequency and time gaps
   ↳ Sample feature extraction - tweets counts, average time between tweets, % tweets with all CAPS, emoji counts, emotionally charged words count, tweet length, hashtag frequency
   - this should result in feature matrix!
2. Feature Scaling:
   - SKLEARN stuff seen in some academias already
   ↳ KMEANS and DBscan use distance (this has to be done or the number of tweets could dominate emoji counts)

3. Clustering / anomaly detection:
   - KMEANS to use for cluster types
   - DBScan to use for dense clusters + outliers?

4. Cluster tuning with K-MEANS
   ↳ "Elbow method" to see where adding clusters stops improving results
   ↳ "Silhouette score" to evaluate how well separated are the clusters
   "import silhouette_score"

5. Evaluation:
   - silhouette score (KMEANS)
   - visualization of clusters (coaster plot, PCA, heatmap)
   - case studies (2-3 manually inspected users and their results)

6. Interpretation + visualisation:
   - Come up with labels like "stable, escalating, erratic"
   - graphics include - line plots
     - cluster coaster plots
     - timeline heatmaps for tweet activity

- new dataset that looks promising:
  ↳ Israel-Palestine Conflict Tweets dataset
  - over 16 thousand tweets, which could be promising useful
  - the only downside is that there are no usernames, only twitter links → there seemas to be a method called "rehydrate tweets" meaning - fetch the names again using Twitter API (paid) or snscrape
    ↳ apparently SNSCRAPE can be used to build my own dataset by scraping hashtags

- this dataset had some sort of export error, with every date having "Â" symbols next to them, making it impossible for python to recognise it - ended up removing them in excel

- one user had over 25k characters tweet, so I am going to set a character limit to '1000'
- when generating a timeline, one user has one tweet with 10 000 characters, but as it is just one point, it is impossible - it happened on 22.11.23

- key takeaways so far:
  ↳ plots show that not that many people use caps, emojis or hashtags that often
  ↳ not everyone posts all the time, with some users posting about once every 4 days on average
  ↳ average tweet length shows that broader distribution of characters used way above the former

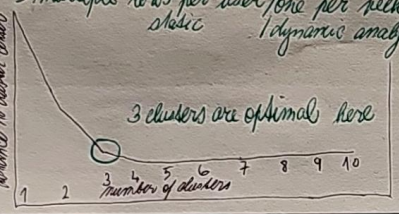
What am I doing?

Analysing Twitter user's ~~behaviour~~ texting behaviour to group them into meaningful categories (I need to figure these out, maybe once I see some experiments) like stable, erratic, or escalating users?

- But what is behaviour?
  - ↳ observable patterns, like often they tweet (frequency)
    - how they write (tweet length, emojis, etc.)
    - what they write (Q4: how is there like an general list of words that I can use?
  - ⇒ behaviour is the combination of frequency, timing and writing style of tweets, that can be quantified into numerical ~~values~~ features

Q1: Do I need to implement a lexicon?
- make your own researched list of intense words and count how many of these would appear per user over time
  - Find books/resources that implemented this already

Q2: How does Elbow method know what to cluster?
- I have to define what gets clustered by the features I extract
- ⇒ when adding more clusters no longer improves how tightly the clusters fit the data. Plot the total error for k = 1 to 10 and look for the "elbow" where the curve bends.
- username | tweet_count | emoji_average | week_1 | weekly_difference |
  - ↳ multiple rows per user (one per week) static / dynamic analysis (changes)



How am I doing it?

I am extracting numeric features from tweets (like frequency, caps, emojis), scaling them, clustering them (Q2: how does elbow method know what to cluster for?) and interpreting the results with visuals

What is my methodology?

Suggested method = tone (how do I define this?) + frequency + keywords (Q3)
- ↳ is tone all the features that I already identified?

Cluster evaluation
- review manually a few users per cluster
- show patterns over time (Q5: is this done manually by me as well?)
- ground truth validation – I don't have any labelled data (Q6) or known behaviours to compare my clusters to, say they are correct, I have to write it as a limitation as this is ~~unlabelled~~ unsupervised data, – is there a way to potentially base this?

END

Q7: Will I have some sort of basic overview visualised? If this was for trust and safety then why am I analysing datasets and not individual users? Would this be a way to handle events happening online?
- I can "detect escalation"
  - ↳ users becoming more intense over each week
  - ↳ if I implement risk, it could be even more emotionally charged words ~~per~~ per week

Q8: calculate the slope (how do I do it?)

Q6: Can I add ground truth validation in this case?
- No, this is only for labelled data, in this case there is no "verified ~~data~~ real-world data"
- maybe if I knew some real public figure that is definitely right/left wing, I could test their tweets to see how the model works?
- What will most likely have to happen is a mention of this in a form of a (disclaimer in the report) as this is an unsupervised system

Q7: What would the final overview have?
- cluster sizes, silhouette score, average values per cluster (data?)
  - example: top 3 users per cluster with a summary of their behaviour
  - ↳ part of this could be a trajectory using plots
- some users will escalate, some de-escalate and some will not change at all
- all of this could be a table ⇒ this poses a question about a need for an interface? Keep this in mind for a/as time

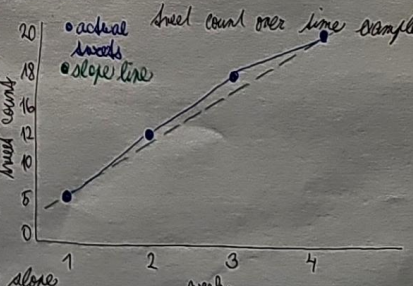
Q3: What exactly is "tone" and am I already capturing it?
- in this sense, tone is emotional intensity, aggressiveness and expressiveness
- besides caps, hashtags and emojis, I am also adding emotion words, which will help determine intensity
- tone = writing style + emotional features

Q4: Am I missing something here?
- as of now, I need to come up with a lexicon of words that will ~~often~~ help determine each cluster (Q1 already touch that)
- definitely some weekly trends
- cluster interpretation – guidance that actually describes each cluster based on its average values
  - ↳ Find academic resources that have done this and come up with your own cluster descriptions

Q5: Showing patterns over time – is it going to be manual?
- this can be both manual and automatic
  - ↳ manual would be picking few users, most likely those with most significant changes and applying plots and describe their trends
  - ↳ automatic – calculate the slope of features per user across multiple weeks
    - ↳ let's say three users per cluster?

Q8: How do I calculate the slope?
- assuming we do this to discover if user's behaviour is increasing, decreasing or staying the same over time?
  - ↳ are they tweeting more over time? Or using more emojis?
    - positive slope = behaviour increases
    - negative slope = behaviour decreases
    - flat slope = behaviour is stable

week	tweet count
1	5
2	8
3	12
4	20



- python calculated the slope as 4.90 tweets per week

What can I do to describe clusters properly
↳ using both KMeans and DBSCAN gives different results that could be aligned as they overlap
↳ possibly the average values of each feature
↳ ↑ tweet count ⇒ spammer
- find features that are high only in one cluster

cluster 0 ⇒ neutral
cluster 1 ⇒ emotional
cluster 2 ⇒ heavy posters
outlier ⇒ all over the place ⇒ spammers, but random

- the cluster shows its prominent features ⇒ average behaviour is a prominent feature for neutral cluster

How do I prove that clustering works?
- lots of users overlap in clusters due to their features

↳ as this is unsupervised, we can't "prove" cluster
↳ only thing that seems logical is similarity with feature behaviour "proving" correct clustering

should I filter each cluster by their highest emotional use or by emoji use? or any other feature, isn't that creating a bias?

↳ if i select features I assume fit in the cluster, that feels very artificial

↳ maybe by selecting "longest timeline" from user in each cluster will show why they are in that particular cluster