Setting up the AI & ML

infrastructure for Hunch

[**Hunch and the Data Points**](#_iode1jmbhwo7) **1**

[**Data Availability**](#_zcqeevsgvjb4) **2**

[**How are polls created & answered today?**](#_jr2xxq95mttq) **2**

[**Features coming soon that will impact data and it’s interpretation**](#_xtma5tk4jwz3) **3**

[User generated content](#_g5643xh8tgzd) 3

[Follower/followee system](#_xpaqupa08aho) 3

[Upvote/Downvote of polls](#_rpc8ioaq5udn) 3

[**What we need help with**](#_8h90ewjpllmh) **3**

[Personalized Recommendation engine](#_7ft69deqmpph) 3

[The content profile](#_ht4hgbau85r) 4

[The user profile](#_9xg3gznbvag1) 4

[The environment profile](#_o1urj1jdzr1s) 4

[Hot Takes](#_hbnwsf4vmbja) 5

[Identification and promotion of viral content](#_iqkwvfj09w7q) 6

[Cross Polls Insights](#_h0o5mulbzz2p) 6

[Search](#_745qfl70ce2b) 6

[Similar polls](#_bglpscawkv9f) 7

[**Resources & Infrastructure Setup**](#_o1zmtcuirejb) **7**

### Hunch and the Data Points

A $25m Series A funded startup, [Hunch](https://hunch.in/) is a pseudonymous social media app where the feed shown to the users is a series of polls only. Users interact with the polls in the form of:

1. View

2. Vote

3. Comment

4. View comments

5. Share

We have data, both from a poll and a user perspective.

Poll data would mean

1. PollId
2. Title
3. Options
4. Poll type - text/image(video in future)
5. Base category
6. Start date
7. Duration
8. Views
9. Votes
10. Comments
11. Share

User data would mean

1. UserId
2. Polls Viewed
3. Polls Voted and option voted for
4. Polls Commented and the comment made
5. Comments liked
6. Replies to comments
7. Polls whose comments were viewed
8. Polls shared

The above is not an exhaustive list of data but shared just for an understanding.

### Data Availability

Hunch has been live with external users starting 26th December and we plan marketing with full force and budget starting 10th January. This means that some data from users has started flowing in and we can expect a more exponential growth in data over the months. Consider that we will have data of 30K monthly active users by the end of Feb 23 (which means a much larger set of Top of the Funnel Data)

Like every social media app, interpreting the data is critical and brings about network effects that keep users retained to the app.

### How are polls created & answered today?

1. All polls are being created internally today.
2. A base category is defined for each poll.
3. There are only 20-30 new polls created a day.
4. We have 2 types of feed, Hot and New. The new feed is basically just a chronologically descending list of polls. The Hot feed is generated via a basic ranking algorithm based on weights given to shares, comments, votes, views etc.
5. Both the above feeds are global and not user specific., i.e. if 2 users open the feed they will get the exact same order.
6. Once a poll is answered, the poll is removed from their feed and can be seen on their profile page

### Features coming soon that will impact data and it’s interpretation

#### User generated content

We shall be opening up the feature of creation of polls to other users. We plan to start with influencers and core users and slowly make the functionality available to all. Once this is done, we will have no direct control over the quality and categorisation of content. The number of polls being created may also vary and hence the ranking of polls would be important.

#### Follower/followee system

Users will be allowed to follow other users so that they do not miss out on the comments and polls created by the users they want to follow.

#### Upvote/Downvote of polls

Users will be able to upvote or downvote polls to mark their liking/disliking of a poll or content.

#### Report/Block user

Users will be able to report polls and block users. Reporting of a poll would mark a negative sentiment of the user towards that content/creator. On the other hand, if user A blocks user B, then user A would not be able to see any of user B’s content(poll/comment/profile) and vice versa.

### What we plan to do as 1st steps

We have researched and come up with areas where we need help. Here are a few initial ways in which we want to use data today:

#### Personalized Recommendation engine

This is important both today and even more when UGC and follower system is out -

Recommendation systems rely on 2 key processes: **content based** filtering and **collaborative** filtering.

* A content based filtering will recommend content based on what they like to consume, eg what do they vote on, which topics do they read or make comments on, what do they share. Do they enjoy text/image or video based polls? The system would also lay negative weightage to reported/blocked polls/users.
* A collaborative filtering system will base their recommendations on finding groups of users who enjoy similar content. Say Jane and Tracey’s interests are highly correlated. If Jane votes/shares a poll, a reliable indicator of interest, then the system will also recommend the poll to Tracey.

An ideal system should center around 3 central profiles:

##### The content profile

A poll has a title and description and options, based on that there should be several content tags automatically attached to the poll. Users might self select hashtags and categories but the NLP engine should be primarily responsible for identification and weighing in keywords.

##### The user profile

The user profile is built from the several aspects in the system, eg his search history, his vote/comment/share history, kind of content which the user rejects, the kind of polls he creates, device location, age, gender, behavioral traits(understanding his profile from the way he/she answers),

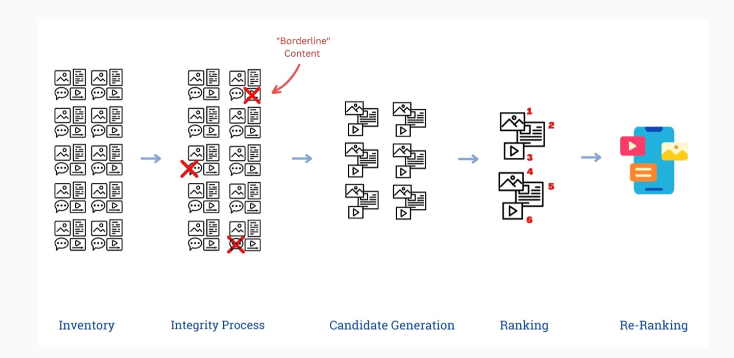
##### The environment profile

Environment profile is based on where the user consumed the content, for eg at work or home or commute as people’s preferences vary given different situations.

The system should compute the strongest statistical match between the content profile, user profile and environment profile that will optimize the user’s interaction with the app and hook him.

Apart from this a user might want to manually follow people to not miss out on any content generated by them. Even if they do not manually follow, the system should automatically recognise the people who’s content a user enjoys seeing and rank that in a user’s personalized feed.

How it generally works



More details on the above diagram can be seen [here](https://www.brookings.edu/techstream/how-do-recommender-systems-work-on-digital-platforms-social-media-recommendation-algorithms/)

As we move further, we would want to consider 2nd and 3rd order profiles as well.

For eg, content profiles of people a user follows, content profiles of people engaging with the same poll as the user (even if there is no explicit follow relationship)

The output can be recommendations of content or other users to follow

#### Hot Takes

Hot Takes is a daily recap that shows you what interesting things happened on hunch yesterday. Hot Takes would be in a ‘Stories’ format that users can swipe up just like reels to see more.

1. Every day this recap will change based on the previous day’s insights gathered
2. This will be made available on the app once a day.
3. Users can see the Hot Takes through the ‘Stories’ feature
4. There would be templates set by us that will be used daily. However, the content varies daily

We want to automate the generation of these hot takes. Currently, the thinking is to generate takes based on the following templates

| **Template Name** |
| --- |
| Trending poll - Polls leading in number of votes |
| Banter poll - Polls leading in number of comments and replies threads |
| Sharable poll - Polls leading in number of shares |
| Comment on fire - Comment which is hot, catching a lot of reaction |
| Vibe earner - profiles who have recently gained a lot of vibe in the system |
| Skewed polls - polls which have a clear majority in an option, can be a segment of category which is skewed eg 80% females think Men should cook etc |
| ‘Close Call’ polls - polls which are mostly neutral |
| Polls where a segment doesn’t follow the majority view - polls where a filter does not follow what the world in general thinks eg, world in general loves normal maggi but 80% of LSR like cheese maggi or 80% men think women should not drive etc. |

#### Identification and promotion of viral content

When we start with UGC, the ranking algorithm will have to be sharpened to identify posts which can attract virality and surface them to the right kind of audience.

Every time a poll would be added to hunch, the title, description and options should be queued up and go through an automated audit. Computer vision should be used to analyze and identify elements in images or text and then categorize the poll with keywords. Any poll suspected of violating the platform’s guidelines should be flagged for human review. Content should also be flagged for duplicate content to prevent plagiarism and maintain originality of content.

After the screening process, the poll should be released to a small pool of a few 100 active users. Metrics such as views, votes, comments, shares, upvotes/downvotes should be analyzed to gauge the video's popularity within its vertical category. Those that will pass through to the next level, where the video is exposed to 1000s of users. Again, the metrics are evaluated, with the top-performing videos passing on to the next level, where they gain exposure to an even larger audience.

The process shouldn’t run entirely on algo. At the higher levels, a person on the content moderation team should manually review the poll to see if the poll does not violate the platform’s term of service or harmful content trying to beat the algo.

#### Cross Polls Insights

We want to play around with data and create fun facts which are not that obvious. For eg, 80% of nonvegetarian people who eat outside 3 times a week. Now this result can come from analyzing and finding correlation across several polls. 1 poll which asks are you a vegetarian/not and then the second asking how many times do you eat outside.

#### Search

We will soon have a search functionality on hunch, where users can search for any text and the results not just directly having those words but also correlated to it are shown.

#### Similar polls

We have a web page where people can see the polls that are shared by their friends. Now, this page has a section for recommendations or polls you may also be interested in, which should ideally be similar or environmentally related to the poll.

#### Resources & Infrastructure Setup

To execute the above requirements, we will need a team. Our current product team has limited knowledge around AI & ML, even though we get support from our sister-concern, Quasar which has know how.

Our current understanding is that we will need:

1. A data scientist who can patternize our data, find the right formulas, and build quantitative models that can be applied.
2. An AI/ML engineer who can work with the data scientist to build the model using technology frameworks like Python
3. A senior consultant who can oversee and guide on the process of setting up the team & corresponding infrastructure so we don’t mess up.

#### Leveraging ChatGPT

We’ve seen how useful ChatGPT is already with prompts that are designed to create meaningful questions itself. We would like to assess how to use ChatGPT to build AI generated content (AIGC) that can be served to the users as regular feed.

### Submission

Create a project proposal for implementing the [Personalized Recommendation engine](#_7ft69deqmpph)

1. You can be creative with the scope of things to include in the report but below are a few suggestions that can guide you to think about the report.
   1. Problem you are trying to solve.
   2. Different models to solve this problem.
   3. Creating, training and testing the model you chose and it’s POC
      1. Which of these models are you going to use to solve this problem?
      2. Why will you use it this way?
      3. What are the constraints of this model?
      4. What are the primary and the secondary goals?
   4. Execution on live data
      1. Building the Architecture to deploy this model in production.
      2. How would you set up the architecture for this model?
      3. What is the technology used to deploy this model?
   5. Research cycles and improvements
      1. How will you plan the research cycles?
      2. How will you decide when improvements will be done and implemented in production?
      3. What are the thresholds for the same?
   6. **What will be a good POC for this model and its execution on live data?**
   7. Scaling
      1. Limitations of the model when entering big data.
      2. Will your system be able to handle it if we have millions of users generating TBs of data?
      3. Parallel processing of data.
2. Potential questions you can ask yourself while writing this report.
   1. If you would be implementing such a model and it’s underlying infrastructure, how would you implement it?
   2. What technologies would be used?
   3. How would you set the timeline to execute it?
   4. How would you structure this architecture?
   5. What are your past experiences, if any, of implementing something like this?
   6. What kind of resources would you require in order to execute it?
   7. And most importantly, why would you do it the way you propose to do it?
   8. Feel free to be opinionated about different models and architectures.
3. The more summarized and shorter the report, the better it is.
   1. The limit for the report is 9 pages.
   2. More points for written communication that conveys a lot in lesser words through data, flowchart and pictures.