Mindhive

AI-based Expertise Matchmaking and Insight Generation Algorithms

Research Progress Whitepaper — July 2021

AI-based Expertise Matchmaking
and Insight Generation Algorithms

Research Progress – July 2021
Mindhive.org

Front and back cover images by Giorgio Grani
https://unsplash.com/photos/cDsakT8anSw

Contents

Problem Statement

Mindhive identified a gap in the market for a faster, more streamlined, and scalable crowdsourcing network solution, offering a network of innovative consulting minds to provide tools for rapid insight and innovation in a manner not previously achieved anywhere in the world.

To achieve a state-of-the-art crowdsourcing policy development platform, Mindhive initiates two research priorities incorporating artificial intelligence (AI). Two research products will be presented in this paper. Depicted in Figure 1, these priorities are two AI-based algorithms to support a more engaging Mindhive platform, namely expertise matchmaking as well as insight summarisation.

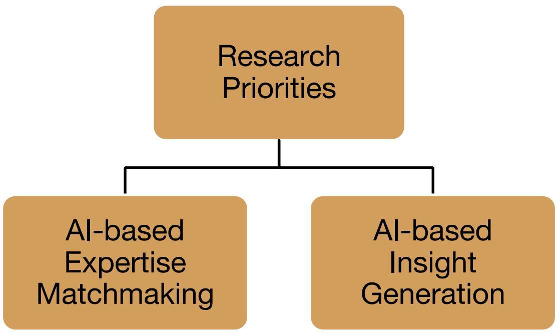

Figure 1. Mindhive's Research Priorities

Each of these two research priorities will be undertaking similar workflow as depicted in Figure 2. Firstly, exploratory data analysis will be done for each task. After that, we could do data preprocessing to clean up the data from personally identifiable information (PII), as well as removing unhelpful signals that could become noise, such as the letter case as and stop words removal on some scenarios. The next steps are training and testing phases to tune the model parameters. The last step is the implementation, which includes deploying the research outcomes into production level, as well as conducting the user acceptance test to evaluate the helpfulness and correctness of the algorithms.

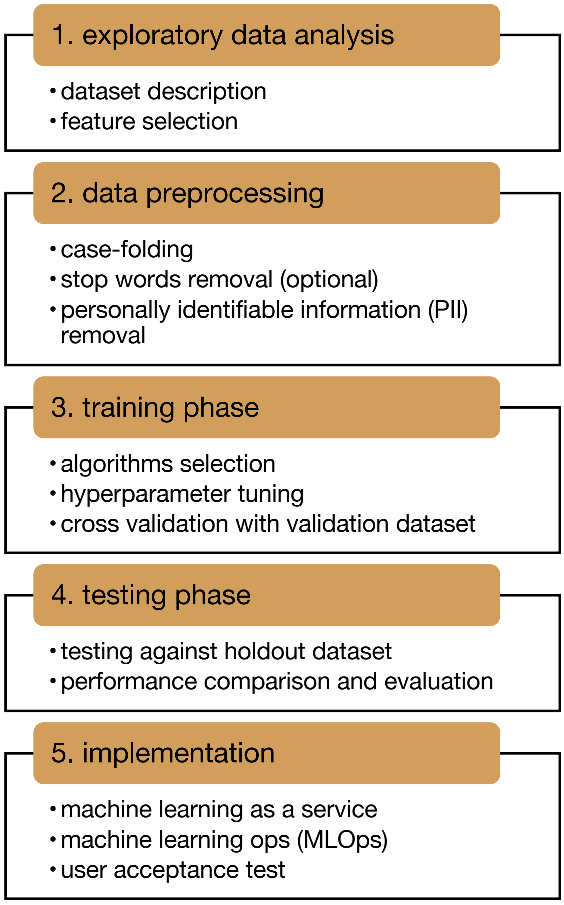

Figure 2. Default research cycle for each machine learning idea

The resources of this research are available by invitation[1]. The sections below describe the two products in more details.

User Matchmaking

Mindhive is specifically developing a novel algorithm, termed Wildcard algorithm, which identifies and connects individuals who show high probability of meaningfully contributing (insight, perspective) to unrelated problems i.e., not matched based on collected or meta data.

The algorithm identifies community members who's in-platform network, discussion input and discussion

[1] In the interest of promoting reproducibility in machine learning research, the source code can be accessed at https://bitbucket.org/mindhivedev/mindhive-r-d by invitation.

interactions would help facilitate, seed, or antagonise further conversation – in turn, creating an environment from which deep insight is surfaced. It also allows for the accurate prediction of groups of individuals whose interaction synergy leads to greater and deeper insight generation.

The effectiveness of the new algorithm will be demonstrated through the effective connection of members to unrelated problems that result in solutions that would not have otherwise been found within a predicted community i.e., a tattoo-artist solving an oil spill problem.

The algorithm is scalable and applicable to any match-making industry where alignment of unconnected pairings provides a competitive or synergistic advantage e.g., dating, recruitment, specialised services (team composition in armed forces or protective services, creative pairings of creative directors and copywriters, teams for scientific research or social policy).

Given recent advances in natural language processing, it has the potential to significantly contribute to the Wildcard algorithmic understanding of how members from different language, education, and cultural backgrounds to unfamiliar problem spaces – or vice versa can be connected, by translating the problem question through the cultural lens for a different perspective.

The implementation of the expertise matchmaking module will have business impacts on increasing user engagement as well as user retention. On a discussion-basis, the implementation of this research would also increase the quality and quantity of posts and comments within discussions.

Figure 3 illustrates the sample implementation of the Wildcard algorithm during discussion creation within Mindhive platform. The algorithm would return select few users who have met the criteria and ranking, and the discussion creator could invite these suggested users to participate in the discussion.

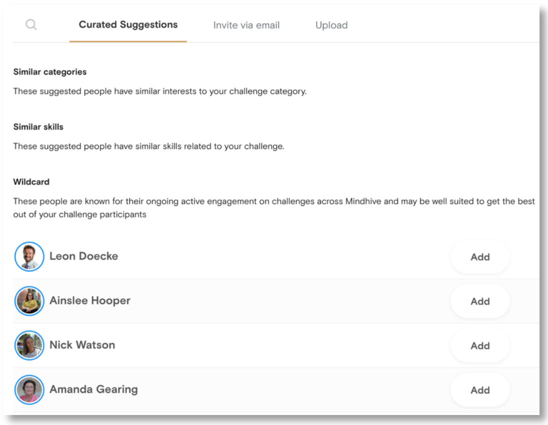

Figure 3. Matchmaking algorithm

In terms of the technical perspective, the expertise matchmaking consists of two subproblems within it. Depicted in Figure 4 are the two subproblems within Expertise Matchmaking problem, namely the matchmaking algorithm itself, and then followed by ranking algorithm.

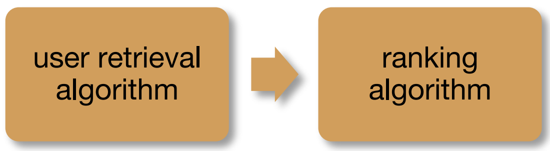

Figure 4. Subproblems within Expertise Matchmaking

There are several features available that might be helpful to consider on retrieving and ranking users on the matchmaking algorithms.

1. Rewards and recognition
2. Skills and interests
3. Recent user contribution activity
4. User gained engagements (number of likes and comments)
5. Recent user consumption activity

Insight Generation

Figure 5 shows the sample of highlighted texts or insights. The idea is to automatically highlight important key takeaways (if any).

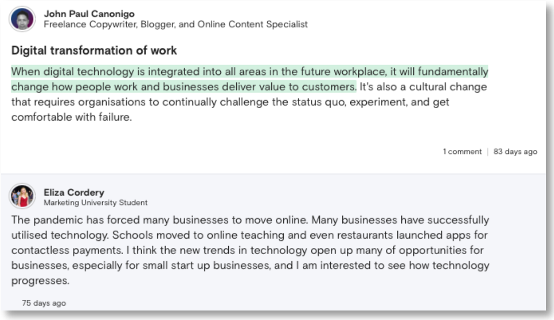

Figure 5. Mindhive's discussion users could highlight users' contribution to attract attention we call it as insights

Figure 6 shows that Mindhive users could categorise the highlighted texts or insights into topics. The insight itself is the verbatim highlighted passage from users' contribution in a discussion.

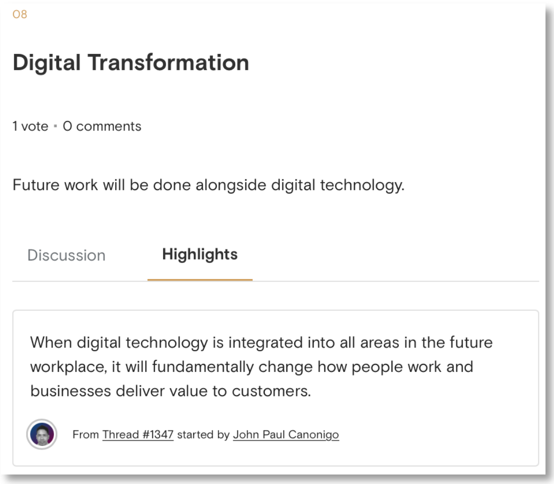

Figure 6. Image from a Mindhive challenge written by John Paul Canonigo [2]. Insights lie under the Highlights tab and could be categorised into topic

In terms of the technical perspective, the insight generation formulation could be stated as the following. Given a text of a post, find the sentences that could be considered as important (if exist) and could also be considered as a new insight relative to other posts within discussion. Each sentence or even each word in the document will be labelled either 0 or 1, where the symbol 1 signals that part is important and should be highlighted, and symbol 0 if it is considered to be not important.

This could be considered as text summarisation, particularly the extractive summarisation where it extracts key sentences in verbatim. This is to differentiate with abstractive summarisation where it tries to reproduce the key important parts of the corpus by paraphrasing them (Carenini, Chi, & Cheung, 2006).

Related previous studies including, but not limited to TextRank (Mihalcea & Tarau, 2004) and Lexrank (Erkan & Radev, 2004) which uses graph-based approaches unsupervised learning, and (Tang, 2019) which used neural networks-based approaches. .

Datasets

The following sections will describe the datasets that could be used for insight generation and matchmaking algorithms.

Insight Generation

The hierarchy of a discussion within Mindhive platform starts from a *challenge*. Challenge is a discussion or a question posted by a user, followed by descriptions, photos, resource links, tags, categories, and other features if it need be. It is worth noting that not all challenges have all of these features, as these supporting information are optional. A challenge could be open to public or private. It could also be a deleted or a drafted challenge. Some preprocessing should be executed to exclude some of these data. Under a challenge, other users could participate by writing *posts*. The post could be commented or liked by users.

Figure 7 illustrates the lengths of questions asked by the challenges' initiators. Most of the questions are considered to be short

[2] https://mindhive.org/challenges/1561/how-do-you-envision-the-future-of-work-in-the-post-covid-era/ideation?modal=highlight-category&detailsId=2029&number=8

texts, with majority of them having less than 15 words.

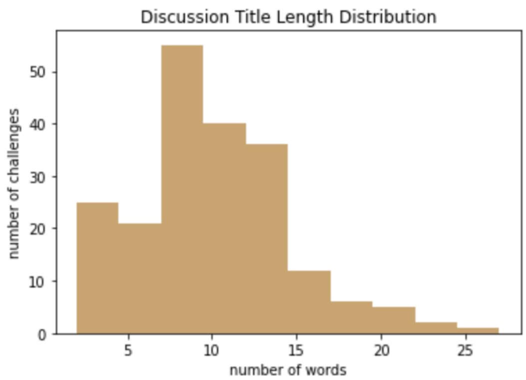

Figure 7. Discussion title length distribution

Figure 8 illustrates the variation of discussion description lengths. This feature could accompany discussion title as the primary textual feature. However, it is worth noting some discussion within the dataset could have empty descriptions. The other features that could accompany these textual data are tags and categories of the discussions. In terms of expertise matchmaking, we also have the data on the users who contributes to the discussions, along with larger audience of users who have seen the discussions.

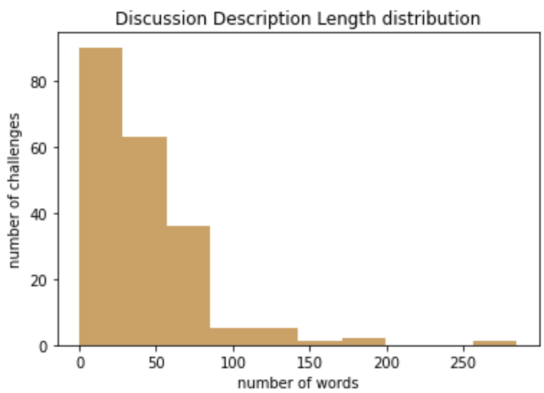

Figure 8. Discussion description length distribution

Depicted in Figure 9 are the top words within the challenge dataset. Some of these top keywords include *Australia*, *people*, *business*, *world*, along with other words.

Figure 9. Top words on Challenges dataset

Expertise Matchmaking

Apart from the dataset that is based around *discussion*, there is also another potential to create another dataset based on the user-level. A *user* on Mindhive platform has their basic PIIs, such as first name and last name, and also interests-, skills- and occupations- related information. This is particularly useful on the expertise matchmaking problem, where the self-reported user profiles could be used as the one of many determining criterion to match users with challenges that they could be interested to contribute.

Figure 10 illustrates the relationship between users and interests that has at least five users self-reporting their interests. Each yellow dot indicates the presence of relationship between a User and an Interest object.

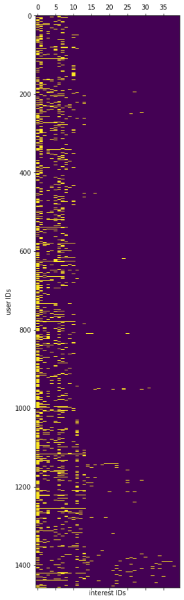

Figure 10. Binary matrix showcasing the relationship between the "Users" and every "Interest" object

User Engagement

Understanding user engagement and loyalty with the product would be helpful to define the feasibility of the Wildcard expertise matchmaking project.

We defined four tiers of platform engagement on a user-level basis within Table 1, which includes:

Level 0 – Registration. Users show general interests to use the platform by completing the registration process into the platform.

Level 1 – Consumption. Users are attracted to consume contents offered by Mindhive platform. This data is currently not available. However, an effort will be implemented in the Mindhive platform to store users' consumption activities.

Level 2 – Passive participation. Users interact albeit passively, such as sharing discussions links, or liking posts.

Level 3 – Active participation. Users show interests to contribute to the discussions by posting their thoughts, commenting in posts, as well as highlighting ideas and insights from posts.

Level 4 – Initiating discussions. Users create discussions in Mindhive platform.

Table 1. Mindhive platform engagement on a user-level basis

Engagement Level		Unique users
Level 0 - Registration		11,427
	signs up	11,427
	verifies the email address	11,045
Level 1 - Consumption		N/A
	views a challenge	N/A
Level 2 - Passive Participation		1,788
	shares a challenge link	N/A
	joins a challenge	1,592
	likes a post or likes a comment	275
Level 3 - Active Participation		389
	posts in a challenge	298
	highlights a challenge	44
	comments in a post	252
Level 3 – Discussion initiator		322
	writes a challenge	322

Additional Datasets

We have encountered the issue of insufficient volumes of the primary datasets, which includes the numbers of discussions currently available, as well as the user base numbers as depicted in Table 1. The unhandled issue of insufficient training dataset could lead to several issues, which include infeasibility of using deep learning-based algorithms as it needs high volume of data, as well as the potential issue of overfitting or high bias during training.

That being said, some secondary datasets outside of the Mindhive platform would be necessary to suffice the requirements of the data-hungry deep learning algorithms (Marcus, 2018). One alternative solution is to use other similar datasets such as abstractive dataset from (Hermann et al., 2015) on insight generation problem. This dataset consists of two corpora, collected from CNN and Daily Mail websites. These two corpora include human-annotated bullet point summaries contained in the article. Another possible additional dataset is adaptation version of abstractive summarisation dataset produced by (Nallapati, Zhou, dos Santos, Gulçehre, & Xiang, 2016), as used by (Tang, 2019).

Another possible scenario is considering alternatives algorithms[3], such as utilising unsupervised learning algorithms, or even exploring the possibility of using transfer learning methods (Zhong, Liu, Wang, Qiu, & Huang, 2019). Baseline results for insight generation could also be obtained by using third-parties libraries, such as the one provided Huggingface library, with the MultiNLI dataset (Williams, Nangia, & Bowman, 2018).

Challenges

There are some challenges which need to be addressed within expertise matchmaking as well as insight generation The following sections describe some of the potential issues related to responsible AI (Google, 2021), particularly in the context of fairness (Satell & Abdel-Magied, 2020) and neutrality.

Maintaining Neutrality

With regards to insight generation algorithm which automatically highlights the important subtexts (if any) on users' contents, it is important that Mindhive remains neutral and does not take any side of discussion polarity.

Some discussions on Mindhive creates strong stance towards the topics, creating in strong polarity between the participants and the readers. For instance, discussions about a policy coined by a politician named John Doe that sparks debate. How do we make sure that Mindhive as the discussion platform remain in neutral? Would implementing the automatic highlighter hurt this neutrality? How do we make sure that the AI algorithms able to accommodate ideas from people who are pros, cons, and none in this situation?

With this consideration in mind, human involvement might be the first step as a precautionary act to prevent issue like this from happening. The moderators or administrators of the discussions would have to approve the automatic highlighter as opposed to fully automatic process.

Inviting All Stances

On Wildcard expertise matchmaking, the similar issue related to fairness might also arise. How do we make sure that the suggested users returned by the matchmaking algorithm represent all polarities within a topic? How do we make sure that the algorithm will not only invite people from a certain wing and neglecting the others?

Remarks

The following are interim key takeaways regarding our ongoing research projects.

1. **The quantity of discussions and user profiles dataset must be increased** to satisfy the requirements of leveraging deep learning algorithms.
2. However, it is important to note that exponentially expanding the volume of datasets organically is not feasible in most of the cases. That being said, we should be **focusing our efforts to leverage more creative ways to work around the dataset volume challenge**. One could use publicly available datasets to accompany the primary dataset. This includes but not limited to external datasets from previous relevant studies (Hermann et al., 2015), or even scraping publicly available contents from the world wide web, such as Medium articles along with

[3] https://paperswithcode.com/task/extractive-document-summarization

their highlighted key points for our insight generation project.

3. Additional features would be needed to perform expertise match making. To suggest potential contributors to discussions, the heuristic option is to match the skillsets and self-reported information such as education and occupation. To expand the suggested users beyond these scopes, we could **store activities within the platforms**, such as reading consumption behaviours. That way, we could expand users' information beyond the self-reported profiles.
4. Another suggestion that we could do is **to leverage transfer learning algorithms**, which has been learned from other separate datasets. With little to no hyperparameter tuning, we could evaluate the performance by applying the algorithms into our primary datasets.

Sources

Carenini, G., Chi, J., & Cheung, K. (2006). Extractive vs . NLG-based Abstractive Summarization of Evaluative Text : The Effect of Corpus Controversiality. *INLG '08: Proceedings of the Fifth International Natural Language Generation Conference*, 33–41. Retrieved from http://aclweb.org/anthology-new/W/W08/W08-1106.pdf

Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*, 457–479.

Google. (2021). Responsible AI Practices. Retrieved from https://ai.google/responsibilities/responsible-ai-practices/

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. *Advances in Neural Information Processing Systems*, *2015-Janua*, 1693–1701.

Marcus, G. (2018). Deep learning: A critical appraisal. *ArXiv Preprint ArXiv:1801.00631*.

Mihalcea, R., & Tarau, P. (2004). {T}ext{R}ank: Bringing Order into Text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404–411. Retrieved from https://www.aclweb.org/anthology/W04-3252

Nallapati, R., Zhou, B., dos Santos, C., Gulçehre, Ç., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. *CoNLL 2016 - 20th SIGNLL Conference on Computational Natural Language Learning, Proceedings*, 280–290. https://doi.org/10.18653/v1/k16-1028

Satell, G., & Abdel-Magied, Y. (2020). AI Fairness Isn't Just an Ethical Issue. Retrieved from Harvard Business Review website: https://hbr.org/2020/10/ai-fairness-isnt-just-an-ethical-issue

Tang, J. (2019). *Key - sentence extraction with Neural Network*.

Williams, A., Nangia, N., & Bowman, S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 1112–1122. https://doi.org/10.18653/v1/n18-1101

Zhong, M., Liu, P., Wang, D., Qiu, X., & Huang, X. (2019). Searching for effective neural extractive summarization: What works and what's next. *ArXiv Preprint ArXiv:1907.03491*.

Mindhive