Motivation

Users come to Facebook for many reasons—to be inspired, entertained, connected, etc. We need content understanding through affective response (AR), rather than merely topic classification.

Contributions

1. Designed novel AR taxonomy.
2. Collected large-scale dataset for AR.
3. Trained a two-tower architecture model.
4. AR model improves recommendation!

Challenge 1: How do we define AR?

Table 1. Taxonomy constructed with UX researchers that is granular enough to cover critical use cases but not tediously long.

| Class      | Definition                                                                 |
|------------|-----------------------------------------------------------------------------|
| Adoring    | Finding something adorable.                                                 |
| Connected  | Feeling more connected.                                                    |
| Good-angered | Constructively angered.                                                      |
| Bad-angered | Toxic/unproductively angered.                                               |
| Amused     | Amused or humourised.                                                       |
| Excited    | Feeling joy or excitement.                                                  |
| Grateful   | Grateful or appreciative.                                                   |
| Informed   | Informed or enlightened.                                                    |
| Inspired   | Motivated or uplifted.                                                      |
| Neutral    | Having a neutral feeling.                                                   |
| Relaxed    | Feeling calm or relieved.                                                   |
| Saddened   | Feeling grief, unhappy, sad.                                                 |
| Scared     | Feeling of concern or fear.                                                 |
| Surprised  | Shocked or astonished (+/-).                                                |
| Touched    | Moved or emotionally stirred.                                                |

Challenge 2: How do we get data?

Three sources: annotation, comments, and engagement.

Human annotation 800k posts with 5 annotators each. Our interrater correlation averaged over 15 classes (0.52) is much higher than that of GoEmotions [1] (0.28), which has 28 classes.

Figure 2. Number of annotations per affect where at least 3/5 annotators agree (3x) or any annotator selects the label (1x).

Challenge 3: How do we model?

Trained a two-tower model for multi-label classification using our data (1M examples from each of the 23 classes).

Figure 4. AR model. The left tower encodes content while the right encodes user information.

Challenge 4: How do we use this model for recommendation?

Offline testing Used the content-tower embedding as a feature in a recommendation model → AUC loss reduction of 8%.

Online testing Two weeks of A/B testing showed integrity violation ↓ (e.g., misinfo, bullying, & harassment) and engagement ↑ (e.g., overall views & positive reactions).

References

[1] Dorottya Demszy et al. GoEmotions: A dataset of fine-grained emotions. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, July 2020.

[2] Jane Dwivedi-Yu and Alon Y. Halevy. The CARE dataset for affective response detection, 2022.