Affective Signals in a Social Media Recommender System

Jane Dwivedi-Yu [1] Yi-Chia Wang [1] Lijing Qin [1] Cristian Canton Ferrer [1] Alon Y. Halevy [1]

[1]Meta AI

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.

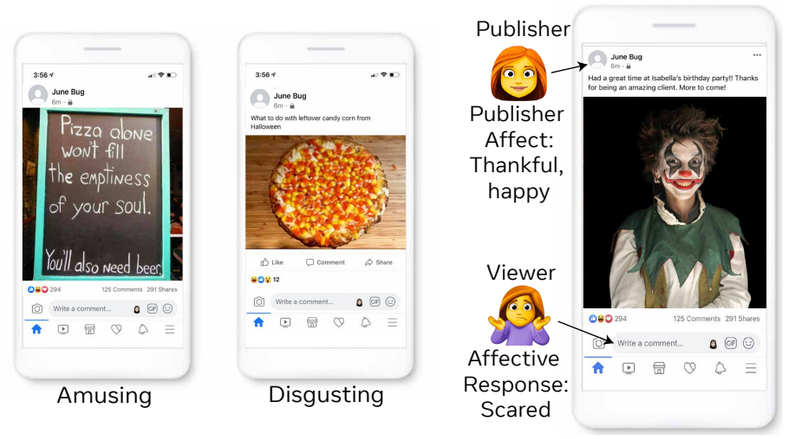

Figure 1. Left: Differing affective responses, same topic (pizza). Right: Publisher affect vs. affective response.

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 humoured.
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.

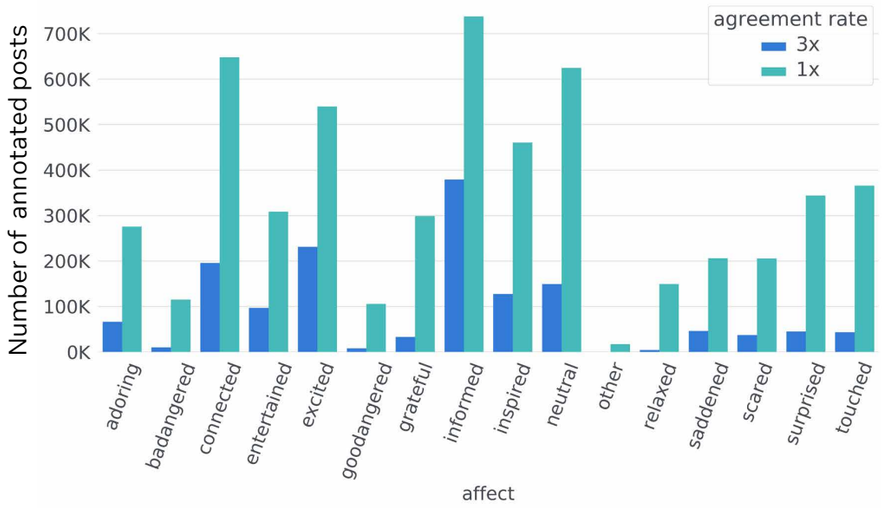

Figure 2. Number of annotations per affect where at least 3/5 annotators agree (3x) or any annotator selects the label (1x).

Comments Labeled posts with the CARE method [2].

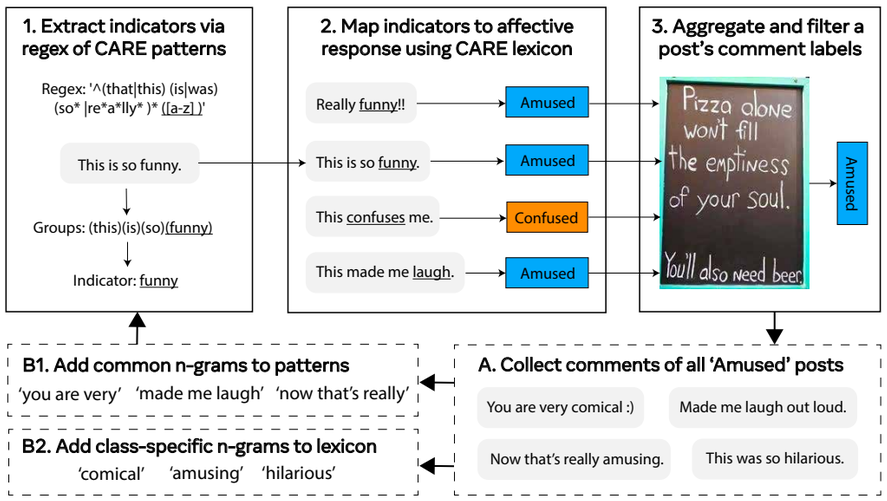

Figure 3. CARE Method. The top part shows the process of labeling a post, while the bottom shows how we expand the patterns and lexicon.

Engagement. Included data from engagement signals: reactions (like, love, wow, etc.), behaviors (e.g., comment, share, click, etc.), and feedback (e.g., reports, hide, skip).

References

[1] Dorottya Demszky 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.

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).

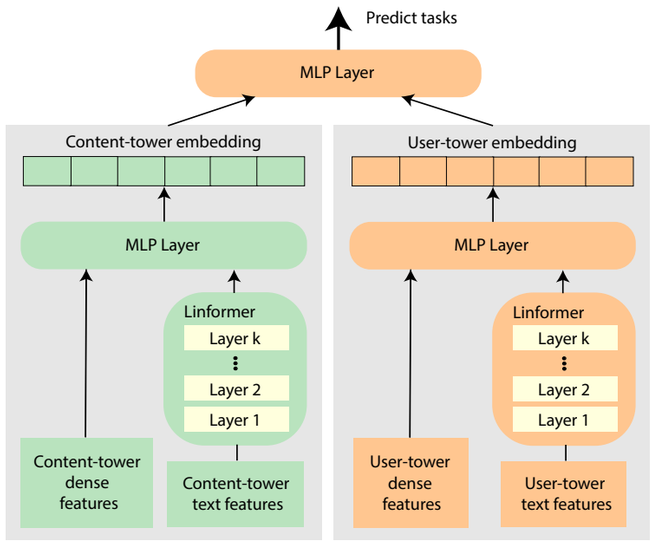

Figure 4. AR model. The left tower encodes content while the right encodes user information.

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).

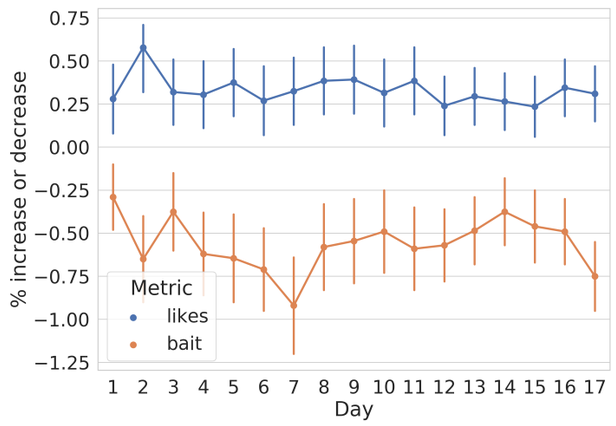

Figure 5. Percent ↑/↓ in overall number of likes and views of engagement bait, respectively.

Deployment! After trends observed in online test continued for two months, our model was deployed at full scale.