Challenges in Providing Automatic Affective Feedback in Instant Messaging Applications

Chieh-Yang Huang [1] **Ting-Hao (Kenneth) Huang** [2] **Lun-Wei Ku** [1]
[1] Academia Sinica, Taipei, Taiwan. appleternity@iis.sinica.edu.tw, lwku@iis.sinica.edu.tw.
[2] Carnegie Mellon University, Pittsburgh, PA, USA. tinghaoh@cs.cmu.edu.

Abstract

Instant messaging is one of the major channels of computer mediated communication. However, humans are known to be very limited in understanding others' emotions via text-based communication. Aiming on introducing emotion sensing technologies to instant messaging, we developed EmotionPush, a system that automatically detects the emotions of the messages end-users received on Facebook Messenger and provides colored cues on their smartphones accordingly. We conducted a deployment study with 20 participants during a time span of two weeks. In this paper, we revealed five challenges, along with examples, that we observed in our study based on both user's feedback and chat logs, including *(i)* the continuum of emotions, *(ii)* multi-user conversations, *(iii)* different dynamics between different users, *(iv)* misclassification of emotions, and *(v)* unconventional content. We believe this discussion will benefit the future exploration of affective computing for instant messaging, and also shed light on research of conversational emotion sensing.

Introduction

Text-based emotion detection has a long-lasting history of research, however, has rarely been used in applications for individual users such as instant messengers. To understand the feasibility of text-based affective computing in the era of mobile devices, we introduced *EmotionPush*[1], a mobile application that automatically detects the emotion of the text message that user received via Facebook Messenger, and provides emotion cues by colors in real-time (Wang et al. 2016). EmotionPush uses 7 colors to represent 7 emotions, which is based on Plutchik's Emotion Wheel color theme (Figure 1 (a).) For instance, when the user receives a message saying "Do you wanna have the brunch?," EmotionPush first classifies this message's emotion as *Joy*, and then pushes a notification on the user's smartphone with a yellow icon (Figure 1 (b)), which is the corresponding color of *Joy*. Later when the user clicks the notification to open Messenger to start the conversation, EmotionPush keeps track on each message that the user receives and uses a color bubble on the top of the screen to continually provide emotion cues

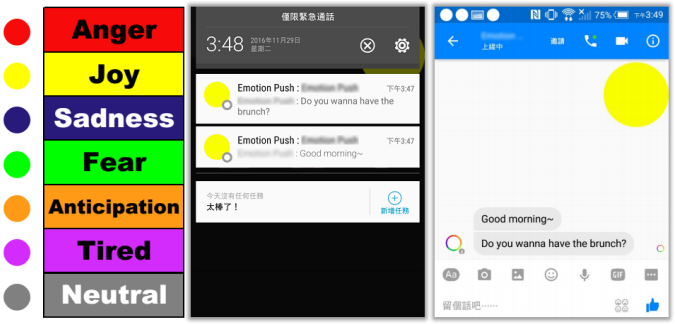

Figure 1: The (a) emotion-color mapping and the system screenshot of (b) colored push notification and (c) colored floating bubble feedback of EmotionPush.

(Figure 1 (c)). The emotion classifiers of EmotionPush were developed by using LibSVM (Fan et al. 2008) and trained on LJ40k (Leshed and Kaye 2006), a dataset contains 1,000 blog posts for each of 40 common emotions. These classifiers were shown to be comparable to the state-of-the-art methods in terms of performance (Wang et al. 2016).

We conducted a deployment study of EmotionPush with 20 participants during a time span of two weeks. 20 English native speakers who identified themselves as Messenger frequent users (ages ranged from 18 to 31 years) were recruited. Participants were asked to install EmotionPush on their personal smartphones, keep it open, and actively use it during the whole study. Each participant was compensated with $20 US dollars. In our study, totally 62,378 messages were recorded during 10,623 conversational sessions, which were automatically segmented with 5-minute user timeouts.

In this paper, we describe the challenges we identified during the deployment study of EmotionPush. While prior work has shown that identifying emotions based on text is possible, we detail the challenges emerged from the deployment of such a system to the real world. Challenges that we identified included modeling the continuum of emotional expressions; referring the detected emotion to the right speaker in a multi-user conversation; considering various expression levels per familiarity between users; handling the classifier errors; and problems derived from nonconventional contents such as long messages, code switching, and emojis.

[1]EmotionPush is available at Google Play: `https://play.google.com/store/apps/details?id=tw.edu.sinica.iis.emotionpush`

Challenge 1: The Continuum of Emotion

EmotionPush uses a *categorical representation* (e.g. *Anger*, *Joy*, etc.) (Klein, Moon, and Picard 2002) of emotions instead of a dimensional representation (valence, arousal) (Sánchez et al. 2006) to reduce users' cognitive load. One natural limitation of applying a categorical representation is the lack of capability of expressing continuum of emotion. For instance, in the following conversation, the user B sent five consecutive messages, which is less likely to express four different emotions, as predicted[2]:

- **A:** Aww thanks!!
- **A:** How's being home?
- **B:** Studying, haha
- **B:** But it doesn't feel like I have been away for one year
- **B:** Nothing has changed here
- **B:** Time is running so slow now
- **B:** And I'm still jetlagged, haha

While prior work explored modeling continuum of information in text, speech (Yeh et al. 2011) and video streaming (Gunes and Schuller 2013), literature had little to say about modeling continuous emotions in a text-based conversation. To the best of our knowledge, none of the existent conversational datasets contain emotion labels, and the continuum property has not been considered in modern emotion labeling systems for conversations. We believe that considering the hidden emotion states to develop the computational models of humans consecutive dynamics of emotion is a promising direction, where a middle-layered computation which captures the nature flow of emotions is necessary.

Challenge 2: Multi-User Conversations

Unexpected challenges were raised by multi-party chatting, which is also known as *Group* or *Channel* in modern messengers. In our study, in which 22.46% of messages were recorded in multi-user chatting groups, we found that providing emotion cues on top of a multi-user conversation would make it difficult for users to concentrate on the running dialog. For instance, in the following conversation between four different users, it is hard to keep track of both the dialog and the emotion of each user at the same time.

- **A:** Oh I'll have it tonight, just can't rsvp on mobile arm
- **A:** *atm
- **B:** ACK
- **B:** I'll mark you down
- **B:** yup, it's tonight :)
- **C:** holy shit this sounds awesome!
- **B:** Artemis is super nerdy rad
- **D:** I want in on this. I'll see if I can make it work with tech
- **D:** I can make it
- **D:** unfortunately my grandparents are coming in tonight so I don't think I'll be able to join :(ha

[2]EmotionPush users do not receive any affective feedback for the messages sent by themselves. In this paper we show colored emotion cues for all messages only for readers' reference. The example conversations will be lightly disguised based on the techniques suggested by Bruckman (Bruckman 2006) on publication.

Furthermore, multi-party conversations also raised challenges in designing user experience . As shown in Figure 1, EmotionPush uses two ways to provide emotion cues: 1) a colored push notification, and 2) a colored bubble that floats on the top layer of the screen. However, both methods were not capable to efficiently convey emotions in multiple-user conversations. While a notification can show the message and its emotion cue simultaneously, it only displays the name of the chat group instead of the name of message sender; users would also find it difficult to identify the corresponding speaker based on bubble's color changes when multiple users are talking. These design challenges of providing affective feedback that considers emotions, texts and users are beyond prior research of online chatting interfaces (Vronay, Smith, and Drucker 1999; Roddy and Epelman-Wang 1998).

Challenge 3: Different Dynamics Between Different Users

Different interaction dynamics occur between people in different context and relationships. One risk of classifying emotions solely based on texts is the neglect of user context, which is known to have strong correlations with user behavior (Baym et al. 2007; Gilbert and Karahalios 2009). Prior work has also shown that language comprehension is more than a process of decoding the literal meaning of a speaker's utterance but making pragmatic inferences that go beyond the linguistic data (Frank and Goodman 2014).

For instance, in our study, we observed that emotion classification worked better on conversations between users who rarely interacted with each other, in which the languages were more formal. The following is an example.

- **A:** Hey man.. hows it going!
- **B:** Hey! It's going well :-)
- **B:** THings are pretty hectic, but I'm trying to get used to the assignments and whatnotn
- **A:** haha sounds like grad school
- **B:** Yup! Haha
- **B:** Weren't you planning a trip to Pittsburgh?
- **A:** I was! But I never ended up coming.. I would still like to but my best bet was recruiting and I asked not to go as there was soem work that came up

On the other hand, the conversations between users who frequently talked with each other often contain informal expressions. The following is an example.

- **A:** Okay I was thinking of getting pierced tomorrow after 6:30? I could theoretically do today between like 4:30-6 but I worry about cutting it too close?
- **B:** I'M DOWN
- **A:** what time would work best 4 u?
- **B:** a little after 6:30 might work better bc of activity fair?
- **A:** Yeah that makes sense!
  [Discussing about inviting other friends]
- **B:** cooooooooooool
- **B:** i can prob get out of helping with teardown haha
- **A:** Its no big if u cant, its open until 9
- **B:** yeeeee

Our observation suggested that EmotionPush could be more helpful for some conversations, in this case, the conversations between people who talked less frequently with

each other, than for others. User context could be helpful to both directly improve emotion classification or identify which conversations EmotionPush can assist better.

Challenge 4: Misclassification of Emotions

Emotion classification is not perfect. It is inevitable that some emotion cues that EmotionPush send are incorrect. For instance, the message of user B in the following conversation should be of *Anticipation* (orange) instead of *Fear* (green).

A: Will it be factory reset, does it have Microsoft office preset
B: Yes, I will factory reset it tonight and if you want, you can have a look at it :)

In the following example, the message of user A should be of *Anticipation* (orange) instead of *Fear* (green), and the message of user B was apparently not of *Sadness* (blue).

A: Hey guys so does 2:30 sound good for volunteering tomorrow? We'll take next week off because of fall break
B: We can leave at 230

Misclassified cases raised the questions that what level of performance is good enough for an realistic application. EmotionPush's classifier achieved an average AUC (the area under the receiver operating characteristic curve) of 0.68, which is comparable to the state-of-the-art performance (Wang et al. 2016). It is noteworthy that humans are not good at identify emotions in text. Prior work showed that humans on average can only correctly predict 63% of emotion labels of articles in LiveJournal (Mishne and others 2005). Our post-study survey also showed that participants did not think the wrongly-predicted emotion colors are harmful to their chatting experiences (average rating = 0.85, ranges from 0 to 4), while they felt the correctly-predicted emotion colors are helpful (average rating = 2.5). Given all these factors, we believe that our emotion classifiers' performances are practical for real-world applications.

In addition to improving emotion classification, challenges also come from designing good user experience around error cases. EmotionPush is good at identifying *Joy*, *Anger*, and *Sadness* (Wang et al. 2016). One potential direction is to use different feedback types (e.g., vibration) to distinguish reliable predictions from uncertain ones.

Challenge 5: Unconventional Content

Similar to most text-processing systems deployed to the real world, EmotionPush faced challenges in handling unconventional content in instant messages. In this section we describe three types of unconventional content we observed in our study: multiple languages, graphic symbols such as emojis, and long messages.

Multiple Languages & Code Switching Real-world users speak various languages. Even though we recruited English native speakers in our study, participants occurred to speak in, or switch to, various languages when talking with friends. For example, user A switched between English and Chinese in the following conversation.

A: How's ur weekend
A: Sorry last night I didn't sleep well and needed to work ..Feel like I'm a zombie today haha
A: 整天腦袋空空的
A: 你們都搬到台北？
B: 哈哈加油喔喔喔
B: 對呀!
B: 淡水附近
A: How r u

Not only text-based emotion classification require sufficient labeled data for training, but also code-switching processing techniques relies heavily on training data (Brooke, Tofiloski, and Taboada 2009; Vilares, Alonso, and Gómez-Rodrıguez 2015). All of these technologies are not capable of processing unseen languages. While prior work explored cross-language acoustic features for emotion recognition in speech (Pell et al. 2009), detecting emotions in arbitrary languages' texts is still infeasible. For deployed systems such as EmotionPush, making design decision around languages it can not understand is inevitable. Currently EmotionPush supports two languages, English and Chinese, but these two modules were developed separately and still can not handle code-switching case such as the example above. In the future, we are looking forward to incorporating a language identifier to provide more concrete feedback (e.g., "Sorry I do not understand French.") to users.

Emoji, Emoticons, and Stickers Graphic symbols such as emojis, emoticons and stickers are widely used in instant messages, often for expressing emotions. For example, the emoticon "¯_(ツ)_/¯" (also known as "smugshrug"), which represents the face and arms of a smiling person with hands raised in a shrugging gesture, was used in the following conversation.

A: when can i come and pick up my Jam and also Goat
B: whenever you want tbh?
B: we're home rn if ur down
B: or tomorrow sometime
B: ¯_(ツ)_/¯

The usages and effects of graphic symbols in on-line chatting have been thoroughly studied (Jibril and Abdullah 2013; Walther and D'Addario 2001; Wang 2015), and techniques of handling emojis in text processing has also been developed (Barbieri, Ronzano, and Saggion 2016; Eisner et al. 2016). However, the current technologies are still not capable to identify emotions from any arbitrary emojis and emoticons, not to mention new graphic symbols are created everyday (e.g., "smugshrug" was just approved as part of Unicode 9.0 in 2016) and stickers are not even text.

Paragraph-like Long Messages Often instant messaging users chunk a long message into smaller pieces and send them consecutively. However, we observed that in our study occasionally users send exceptionally long messages. For instance, a user sent one message that contains 10 sentences (134 words) to warn the former owner of his/her house to clean up as soon as possible, a user sent a 10-sentence message (201 words) to advertise his/her incoming stand-up comedy performance, and a user sent a 9-sentence message (152 words) to discuss an reunion event. In each of these

long messages, the user used multiple sentences to express complex issues or emotions, which made it difficult to conclude the message with one single emotion. While literature showed that emotion classification yielded a better performance on long sentences that contain more words because they bear more information (Calvo and Mac Kim 2013), our observation suggested that long messages that contain many sentences often result in a less-confident or incorrect emotion classification as a whole.

Conclusion & Future Work

In this paper, we describe challenges in deploying an emotion detection system, EmotionPush, for instant messaging applications. These challenges included the continuum of emotions, multi-user conversations, different dynamics between different users, misclassification, and unconventional content. These challenges are not only about providing automatic affective feedback by using text-processing technologies, but also about designing an user experience given the interrelated factors including humanity and languages. Through these discussions, we expect to gain insight into the deployment of applications of affective computing and motivate researchers to elaborate the solutions of tomorrow.

In the future, with the advantage of the developed EmotionPush, we plan to design a mechanism which encourages users to contribute their contents and feedback their emotions to advance this technology where it is most needed.

References

Barbieri, F.; Ronzano, F.; and Saggion, H. 2016. What does this emoji mean? a vector space skip-gram model for twitter emojis. In *Language Resources and Evaluation conference, LREC, Portoroz, Slovenia*.

Baym, N. K.; Zhang, Y. B.; Kunkel, A.; Ledbetter, A.; and Lin, M.-C. 2007. Relational quality and media use in interpersonal relationships. *New Media & Society* 9(5):735–752.

Brooke, J.; Tofiloski, M.; and Taboada, M. 2009. Cross-linguistic sentiment analysis: From english to spanish. In *RANLP*, 50–54.

Bruckman, A. 2006. Teaching students to study online communities ethically. *Journal of Information Ethics* 82.

Calvo, R. A., and Mac Kim, S. 2013. Emotions in text: dimensional and categorical models. *Computational Intelligence* 29(3):527–543.

Eisner, B.; Rocktäschel, T.; Augenstein, I.; Bošnjak, M.; and Riedel, S. 2016. emoji2vec: Learning emoji representations from their description. *arXiv preprint arXiv:1609.08359*.

Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research* 9.

Frank, M. C., and Goodman, N. D. 2014. Inferring word meanings by assuming that speakers are informative. *Cognitive psychology* 75:80–96.

Gilbert, E., and Karahalios, K. 2009. Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 211–220. ACM.

Gunes, H., and Schuller, B. 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing* 31(2):120–136.

Jibril, T. A., and Abdullah, M. H. 2013. Relevance of emoticons in computer-mediated communication contexts: An overview. *Asian Social Science* 9(4):201.

Klein, J.; Moon, Y.; and Picard, R. W. 2002. This computer responds to user frustration:: Theory, design, and results. *Interacting with computers* 14(2):119–140.

Leshed, G., and Kaye, J. 2006. Understanding how bloggers feel: recognizing affect in blog posts. In *CHI'06 extended abstracts on Human factors in computing systems*, 1019–1024. ACM.

Mishne, G., et al. 2005. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access*, volume 19, 321–327. Citeseer.

Pell, M. D.; Monetta, L.; Paulmann, S.; and Kotz, S. A. 2009. Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior* 33(2):107–120.

Roddy, B. J., and Epelman-Wang, H. 1998. Interface issues in text based chat rooms. *ACM SIGCHI Bulletin* 30(2):119–123.

Sánchez, J. A.; Hernández, N. P.; Penagos, J. C.; and Ostróvskaya, Y. 2006. Conveying mood and emotion in instant messaging by using a two-dimensional model for affective states. In *Proceedings of VII Brazilian Symposium on Human Factors in Computing Systems*, IHC '06, 66–72. New York, NY, USA: ACM.

Vilares, D.; Alonso, M. A.; and Gómez-Rodrıguez, C. 2015. Sentiment analysis on monolingual, multilingual and code-switching twitter corpora. In *6TH WORKSHOP ON COMPUTATIONAL APPROACHES TO SUBJECTIVITY, SENTIMENT AND SOCIAL MEDIA ANALYSIS WASSA 2015*, 2.

Vronay, D.; Smith, M.; and Drucker, S. 1999. Alternative interfaces for chat. In *Proceedings of the 12th annual ACM symposium on User interface software and technology*, 19–26. ACM.

Walther, J. B., and D'Addario, K. P. 2001. The impacts of emoticons on message interpretation in computer-mediated communication. *Social science computer review* 19(3):324–347.

Wang, S.-M.; Li, C.-H.; Lo, Y.-C.; Huang, T.-H. K.; and Ku, L.-W. 2016. Sensing emotions in text messages: An application and deployment study of emotionpush. *arXiv preprint arXiv:1610.04758*.

Wang, S. S. 2015. More than words? the effect of line character sticker use on intimacy in the mobile communication environment. *Social Science Computer Review* 0894439315590209.

Yeh, J.-H.; Pao, T.-L.; Lin, C.-Y.; Tsai, Y.-W.; and Chen, Y.-T. 2011. Segment-based emotion recognition from continuous mandarin chinese speech. *Computers in Human Behavior* 27(5):1545–1552.