A New Chatbot for Customer Service on Social Media

Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, Rama Akkiraju
IBM Research - Almaden
San Jose, CA, USA
{anbangxu, liuzh, guoy, vibha.sinha, akkiraju@us.ibm.com}

ABSTRACT
Users are rapidly turning to social media to request and receive customer service; however, a majority of these requests were not addressed timely or even not addressed at all. To overcome the problem, we create a new conversational system to automatically generate responses for users requests on social media. Our system is integrated with state-of-the-art deep learning techniques and is trained by nearly 1M Twitter conversations between users and agents from over 60 brands. The evaluation reveals that over 40% of the requests are emotional, and the system is about as good as human agents in showing empathy to help users cope with emotional situations. Results also show our system outperforms information retrieval system based on both human judgments and an automatic evaluation metric.

Author Keywords
Chatbot; social media; customer service; deep learning.

ACM Classification Keywords
H.5.3 Information Interfaces and Presentation: Group and Organization Interfaces.

INTRODUCTION
Social media has changed the way users approach customer service. Nearly half of U.S. Internet users are turning to social media for help, as they can easily send off a Tweet or Facebook status rather than call a 1-800 number or draft a detailed email [10]. Twitter users send millions of requests to major U.S. brands monthly. With the rapid increase in the number of user requests, it has become increasingly challenging to process and respond to incoming requests.

To address this challenge, many organizations form dedicated customer service teams responding to user requests on social media. The team consists of dozens or even hundreds of human agents trained to address users' various needs [9]. However, manually addressing requests is time-consuming and often fails users' expectations. Recent studies show that 72% of users who contact a brand on Twitter expect a response within an hour [19]. Yet, our analysis of 1M conversations shows the average response time is 6.5 hours. This gap motivated us to explore the feasibility of chatbots for customer service on social media.

There has been a long history of chatbots powered by various techniques such as information retrieval and template rules [15]. Deep learning techniques have been recently applied to natural language generation; however, prior work focuses on general scenarios without specific contexts [7]. Lessons could also be informed by studies of social Q&A [5, 6, 13], where users may ask *informational* questions about products or services. Yet, it is not clear how such question types can be applied for customer service.

In this work, we create a new conversational system for customer service on social media. State-of-the-art deep learning techniques such as long short-term memory (LSTM) networks are first applied to generate responses for customer-service requests on social media. The system takes a request as the input, computes its vector representations, feeds it to LSTM, and then outputs the response. The system was trained on nearly 1M Twitter conversations between users and agents from 60+ brands.

In the evaluation, we conduct a content analysis revealing two major themes related to user requests on social media: *emotional* and *informational*. More than 40% of the requests are *emotional* without specific *informational* intents. Our system performs nearly as well as human agents in providing empathy to address users' *emotional* requests. In addition, we find that our system received significantly higher ratings than information retrieval (IR) system in both human judgments and an automatic metric.

CUSTOMER SERVICE CHATBOT VIA DEEP LEARNING
The conversation between users and customer service agents on social media can be viewed as mapping one sequence of words representing the request to another sequence of words representing the response (see Figure 1). Deep learning techniques can be applied to learn the mapping from sequences to sequences [17].

Sequence-to-Sequence Learning
The core of the system consists of two LSTM neural networks: one as an encoder that maps a variable-length input sequence to a fixed-length vector, and the other as a

CHI 2017, May 06-11, 2017, Denver, CO, USA
© 2017 ACM. ISBN 978-1-4503-4655-9/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3025453.3025496

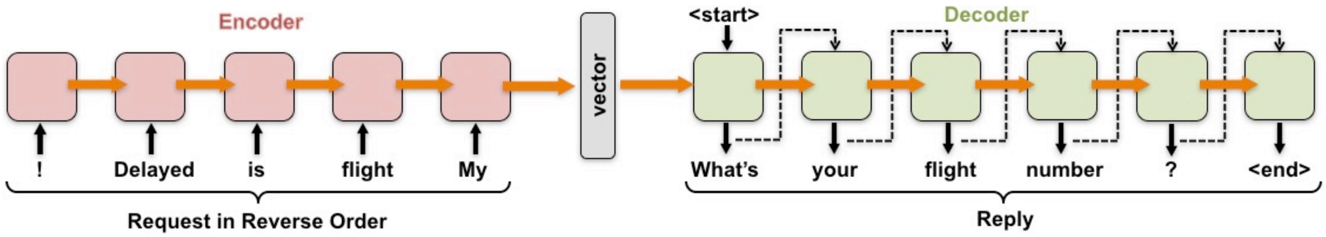

Figure 1. Sequence-to-sequence learning with LSTM neural networks.

decoder that maps the vector to a variable-length output sequence (Figure 1). The advantage of LSTM is that it can store sequential information over extended time intervals and learn to block or pass on information depending on its importance. Following [17], the encoder LSTM reads each input sequence in reverse (Figure 1). This helps the learning algorithm establish a connection between two sequences.

Word Embedding

Words in a user's request cannot be directly used as inputs for LSTMs; each word needs to be converted to a feature vector. Traditional lexicon-based methods [12] can convert words into feature vectors, and many words from social media don't exist in current lexicons [4]. Other feature representations such as n-grams treat words as discrete elements, which would result in a high dimensional vector and, accordingly, a large number of parameters have to be learned. This may cause data sparsity when the amount of training data is incomparable to the number of parameters.

Our system adopts a word embedding method, word2vec neural network language model [8], to learn distributed representations of words from customer service conversations in an unsupervised fashion. The idea of word2vec is that each dimension of the embedding represents a latent feature of the word, which can capture useful syntactic and semantic properties. For example, in a discrete space, words such as "sorry", "apologize", and "glad" are equally distant from each other; but word2vec can represent these words in a continuous space and the distance between "sorry" and "apologize" is shorter than the distance between "sorry" and "glad".

Implementation

62 brands were selected according to three criteria. 1) A brand has a Twitter account dedicated to customer service (e.g. ATTCares). 2) A large variety of brands is covered to enhance the generalizability of our findings across product categories. 3) National brands are selected so that a national sample from crowdsourcing is suitable for evaluation tasks.

The conversation data was collected by the Twitter public API. We used the Streaming API to capture tweets that @mention any of the brands; we also continuously collected the most recent tweets from each brand. We next matched each reply with its request based on the "in_reply_to_status_id" and "in_reply_to_user_id" fields, and thus reconstructed the conversation. Since the Streaming API only contains a sample of user tweets, we also used the Search API to get additional tweets, which were appeared in the "in_reply_to_status_id" field, but were not captured by the Streaming API.

Over 2.6M user requests were collected and only 40.4% of them received replies. 87.6% of the conversations only have one turn (one user request with one agent reply). The collected conversations happened between Jun. 1 and Aug. 1, 2016. 30K of the 1M conversations were stratified sampled from the brands for evaluation and the rest were used to develop our system. Several steps were performed to create the system:

Step 1: Clean the data. We removed non-English requests and requests with images. All the @mentions were also removed in the training and testing data.

Step 2: Tokenize the data. We built a vocabulary of the most frequent 100K words in the conversations.

Step 3: Generate word-embedding features. We used the collected corpus to train word2vec models. Each word in the vocabulary was represented as a 640-dimension vector.

Step 4: Train LSTM networks. The input and output of LSTMs are vector representations of word sequences, with one word encoded or decoded at a time. In view of the clear advantage of deep LSTMs over shallow LSTMs in reported sequence-to-sequence tasks [17], we trained deep LSTMs jointly with 5 layers x 640 memory cells using stochastic gradient descent and gradient clipping.

EVALUATION

We conducted a content analysis to identify themes related to user requests on social media, and examined how the system performs in responding to requests with different themes. The system was compared with actual human agents as well as a standard information retrieval baseline [15], where we retrieved the response whose associated request is most similar to a new request. The similarity measure was based on a TF-IDF weighted vector space model implemented in Apache Lucene [20]. The quality of the generated responses was measured by human judgments and an automatic evaluation metric.

Content Analysis

Following qualitative analysis methods [16], two hundred requests were sampled and coded using a bottom-up approach. The requests were first segmented into the smallest logical units. A first pass was then performed to assign categories to the units and subsequent passes were made to revise and aggregate the categories. We found that there were two types of request:

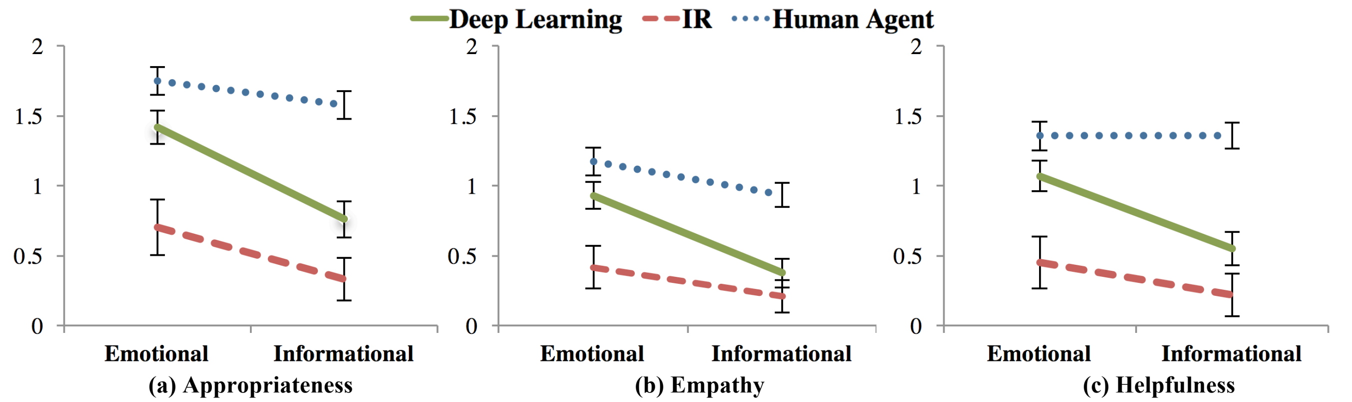

Figure 2. Comparison of human ratings on three dimensions by agent and request types. The two-way ANOVA results for the interactions between agent and request types are (a) $F_{(2, 594)} = 5.61$, $p < .001$; (b) $F_{(2, 594)} = 3.48$, $p < .05$; (c) $F_{(1, 594)} = 7.18$, $p < .001$.

1) *Emotional Request*. In *emotional* requests, users intend to express their emotions, attitudes or opinions toward a brand without explicitly seeking specific solutions (see examples in Table 1). 2) *Informational Request*. Requests are sent with the intent of getting information that may help users solve their problems. This request type is similar to *informational* question identified in social Q&A sites [5].

We recruited two annotators to code another sample of 200 requests using the taxonomy. First, the coders received training in which they were introduced to the themes, definitions, and examples. They then coded requests on a smaller sample of the data and resolved disagreements. Then, they independently coded the requests. Agreement between the coder was high (kappa coefficient = 0.79, $p < .001$). After disagreement was solved, 40.5% of the requests were *emotional* and 59.5% of them were *informational*.

Human Evaluation

Three evaluation measures were derived from prior work to assess the response quality: 1) *Appropriateness*. An appropriate response should be on the same topic as the request, and should also "make sense" in response to it [15]. 2) *Empathy.* The reply should give individualized attention to a user and make s/he feel valued [14]. 3) *Helpfulness.* A helpful reply should contain useful and concrete advice that can address the user request [6].

Emotional	Informational
User: Your customer service is horrible! You don't even know how to deliver packages! *Agent*: I'm sorry to hear that. What was the delivery date provided in your order?	*User*: How do I disable the alarm system on my 2009 Toyota matrix?! Thanks in advance! stranded work in the am gotta get there. *Agent*: Hi, you may visit: (URL)
User: Comcast is the worst. Period. *Agent*: Hi, please let us know if you need assistance.	*User*: Downloaded Nike training club app but won't let me in. Just shows the loading circle before the error message occurs *Agent*: We're here to help. Please dm us the device you're using and we'll get started.
User: Can't wait to travel with you next week for the 1st time of my life with you ;) *Agent*: We can't wait to see you! We're excited to have you onboard with us soon!	*User*: I'm booking a flight and I am not seeing the 25% back on my rewards. Do I have to book on PC? *Agent*: Hi, you can find out more about the pay's website here: (URL)

Table 1. Examples of user requests on social media and their corresponding replies generated by our deep learning system.

Crowdflower was used to recruit participants. All 703 participants were native English speakers and they were 18 or older. The geographic distribution of participants was USA (66.0%), UK (22.8%), Canada (8.5%) and Australia (2.7%). Participants had to fill out at least one gold question in order to participate the survey. 14.1% of participants failed the check and their responses were removed.

In a survey task, participants were first instructed to learn the three rating criteria *appropriateness*, *empathy*, and *helpfulness* with definitions and examples. Then, they were shown a request and asked to rate the three responses from our deep learning system, IR, and human agent respectively. The responses were arranged in random order to control order effects. 200 requests were sampled and thus 600 responses were rated. Each response was rated by 5 participants according to the three criteria. The ratings were made on a 7-point scale from strongly disagree (-3) to strongly agree (+3) with whether the response met the given criterion. Intra-class correlation (ICC(1, k)) of participants' ratings was ranged from 0.60 to 0.87, indicating moderately high reliabilities [2]. The average of participants' ratings of a response was used to measure the quality of the response.

We performed three two-way ANOVA tests to examine the influence of agent type (deep learning, IR, human agent) and request type (emotional, informational) on the three dimensions (Figure 2). On *appropriateness* ratings, the main effects of agent type ($F_{(2, 594)} = 109.46$, $p < .001$) and request type ($F_{(1, 594)} = 44.86$, $p < .001$) and the interaction between them ($F_{(2, 594)} = 5.61$, $p < .01$) were significant at the 0.01 significance level. The two-way ANOVA of *empathy* scores revealed the significant main effects of agent type ($F_{(2, 594)} = 52.37$, $p < .001$) and request type ($F_{(1, 694)} = 44.21$, $p < .001$), and interaction effect between these ($F_{(2, 594)} = 3.48$, $p < .05$). Similarly, the results for *helpfulness* showed the significant main effects of agent and request types ($F_{(2, 594)} = 87.86$, $p < .001$; $F_{(1, 694)} = 25.9$, $p < .001$), and the interaction ($F_{(1, 594)} = 7.17$, $p < .001$).

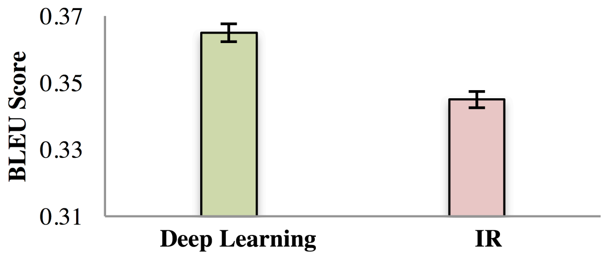

Figure 3. BLEU scores of deep learning and IR systems.

Interestingly, there was no statistically significant difference between deep learning and human agents on *empathy* for *emotional* requests (t-test, $p = 0.15$; Figure 2b), indicating that our system has a similar ability as actual agents to show empathy toward users in emotional situations. Table 1 shows our system recognized different emotional situations and offered empathy accordingly.

Deep learning outperformed IR in all three aspects of ratings (t-test, $p < .01$). The advantage of deep learning over IR was more evident on *emotional* than *informational* questions (Figure 2). However, the performance of both deep learning and IR agents dropped significantly when requests became *informational* (t-test, $p < .001$), Post hoc comparisons indicated that human agent performed equally well on different requests (e.g. t-test, $p = 0.94$; Figure 2c).

Another interesting observation was that, unlike IR, deep learning agent transferred certain writing styles from one brand to another. For example, banking customer service agents often adopted formal language such as "I apologize for the poor user experience" in their responses. However, responses generated by our system became more casual "I'm sorry you feel this way". It is possible that a majority of brands used informal styles on social media. Our system learned these styles and applied them other brands.

Automatic Evaluation

The field of natural language generation has benefited greatly from the existence of an automatic evaluation metric, BLEU [11], which grades an output response according to n-gram matches to the reference (the response from a human agent). We applied this metric to a large testing data set including 30K user requests. Again, deep learning performed significantly better than IR (t-test, $p < .001$; see Figure 3). Moreover, we compared deep learning and IR within each brand. In general, the BLEU scores of deep learning were higher than the scores of IR across brands at the 0.01 significance level.

DISCUSSION AND FUTURE WORK

Traditional customer service often emphasizes users' informational needs [9]; however, we found that over 40% of user requests on Twitter are *emotional* and they are not intended to seek specific information. This reveals a new paradigm of customer service interactions. One explanation is that, compared with calling the 1-800 number or writing an email, social media significantly lowers the cost of participation and allows more users to freely share their experiences with brands. Also, sharing emotions with public is considered as one of the main motivations for using social media [1]. Future studies can examine how *emotional* requests are associated with users' motivation in the context of social media.

Deep learning based system achieved similar performance as human agents in handling *emotional* requests, which represent a significant portion of user requests on social media. This finding opens new possibilities for integrating chatbots with human agents to support customer service on social media. For example, an automated technique can be designed to separate *emotional* and *informational* requests, and thus *emotional* requests can be routed to deep learning chatbots. The response speed can be greatly improved.

Deep learning outperformed IR in all the measures. This is primarily because of deep learning, as a statistical-based approach is much better at handling unseen data and thus more flexible than keyword search approaches. For instance, given a reference reply to the request "my flight is delayed" and one to "my order is cancelled", a deep learning based system is able to generalize the reply in both scenarios and provide meaningful replies to unseen questions such as "my flight is cancelled", for which the most appropriate replies can hardly be retrieved from limited requests/topics available in the training data.

The performance of deep learning and IR systems decreased when requests switched from *emotional* to *informational*, especially in the case of *empathy* ratings. One explanation is that users' informational needs are more diverse than their emotional situations. As a result, it is more challenging to learn and apply the knowledge to *informational* requests. The drop in *empathy* ratings is probably due to the lack of emotional words in *informational* requests. Machine learning techniques are not able to recognize subtle emotions in these requests and response empathetically. Future systems could consider additional contextual information such as users' social media profiles to better understand their emotional status.

We observed that deep learning based system was able to learn writing styles from a brand and transfer them to another. Future work can explore the functionality in a more supervised fashion by filtering the training data with certain styles and specifying the target style for output sentences. This raises new opportunities of developing impression management tools on social media. As written text from brands and individual users affect how they are perceived on social media [18], such a tool can help them create images of themselves they wish to present.

Finally, chatbots on social media offer a new opportunity to provide individualized attention to users at scale and encourage interactions between users and brands, which can not only enhance brand performance but also help users gain social, information and economic benefits [3]. Future studies can be designed to understand how chatbots affect the relationship between users and brands in a long term.

REFERENCES

1. Natalya N. Bazarova, Yoon Hyung Choi, Victoria Schwanda Sosik, Dan Cosley, and Janis Whitlock. Social Sharing of Emotions on Facebook: Channel Differences, Satisfaction, and Replies. In *Proc. of CSCW*, 2015, 154-164.
2. Joan-Isaac Biel, Oya Aran, and Daniel Gatica-Perez. You Are Known by How You Vlog: Personality Impressions and Nonverbal Behavior in Youtube. In *Proc. of ICWSM*, 2011, 446-449.
3. Keith S Coulter, Johanna Gummerus, Veronica Liljander, Emil Weman, and Minna Pihlström. Customer Engagement in a Facebook Brand Community. *Management Research Review*, 2012, *35*, 9, 857-877.
4. Ethan Fast, Binbin Chen, and Michael S. Bernstein. Empath: Understanding Topic Signals in Large-Scale Text. In *Proc. of CHI*, 2016, 4647-4657.
5. F. Maxwell Harper, Daniel Moy, and Joseph A. Konstan. Facts or Friends?: Distinguishing Informational and Conversational Questions in Social Q&A Sites. In *Proc. of CHI*, 2009, 759-768.
6. F. Maxwell Harper, Daphne Raban, Sheizaf Rafaeli, and Joseph A. Konstan. Predictors of Answer Quality in Online Q&A Sites. In *Proc. of CHI*, 2008, 865-874.
7. Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, and Vivek Ramavajjala. Smart Reply: Automated Response Suggestion for Email. In *Proc. of KDD*, 2016, 955-964.
8. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and Their Compositionality. In *Proc. of NIPS*, 2013, 3111-3119.
9. Keith B Murray. A Test of Services Marketing Theory: Consumer Information Acquisition Activities. *The Journal of Marketing*, 1991, 10-25.
10. NIELSEN. State of the Media: Social Media Report. 2011.
11. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proc. of ACL*, 2002, 311-318.
12. James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic Inquiry and Word Count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 2001, *71*, 2001.
13. Tiziano Piccardi, Gregorio Convertino, Massimo Zancanaro, Ji Wang, and Cedric Archambeau. Towards Crowd-Based Customer Service: A Mixed-Initiative Tool for Managing Q&a Sites. In *Proc. of CHI*, 2014, 2725-2734.
14. Leyland F Pitt, Richard T Watson, and C Bruce Kavan. Service Quality: A Measure of Information Systems Effectiveness. *Management Information Systems Quarterly*, 1995, 173-187.
15. Alan Ritter, Colin Cherry, and William B Dolan. Data-Driven Response Generation in Social Media. In *Proc. of EMNLP*, 2011, 583-593.
16. Anselm L Strauss *Qualitative Analysis for Social Scientists*. Cambridge University Press, 1987.
17. Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to Sequence Learning with Neural Networks. In *Proc. of NIPS*, 2014, 3104-3112.
18. Anbang Xu, Haibin Liu, Liang Gou, Rama Akkiraju, Jalal Mahmud, Vibha Sinha, Yuheng Hu, and Mu Qiao. Predicting Perceived Brand Personality with Social Media. In *Proc. of ICWSM*, 2016, 436-445.
19. http://blog.hubspot.com/marketing/twitter-response-time-data.
20. http://lucene.apache.org.