Developers' responses to app review feedback – A study of communication norms in app development

Bastin Tony Roy Savarimuthu[a], Sherlock A. Licorish[a], Manjula Devananda[a], Georgia Greenheld[a], Virginia Dignum[b], Frank Dignum[c]

a - Department of Information Science, University of Otago, Dunedin, New Zealand
b - Faculty of Technology, Policy and Management, TU Delft, The Netherlands
c - Department of Information and Computing Sciences, Utrecht University, The Netherlands
{tony.savarimuthu, sherlock.licorish, manjula.devananda, georgia.greenheld}@otago.ac.nz, M.V.Dignum@tudelft.nl, F.P.M.Dignum@uu.nl

Abstract. Norms are general expectations of behavior in societies. Huge amount of computer-mediated interaction data available in the social media domain provides an opportunity to extract and study communication norms, both to understand their prevalence and to make informed decisions about adopting them. While interactions in social media platforms such as Twitter and Facebook have been widely studied, only recently researchers have started examining app reviews provided by users and the responses provided by developers in the domain of app development. In this vein, a lot of attention has been devoted to study the nature of user reviews, however, little is known about developer responses to such reviews. Additionally, no other prior work has scrutinized the nature of communication norms in this domain. Towards addressing these gaps, this work pursues three objectives using a dataset comprising user reviews and developer responses from Google's top-20 apps used to track running with a total of 24,407 reviews and 2,668 responses. First, based on prior literature in computer-mediated interactions, the study identifies 12 norms in responses provided by developers in three categories (obligation norms, prohibition norms and domain-specific response norms). Second, it scrutinizes the awareness and adoption of these norms. Third, based on the results obtained, this study identifies the need for creating a response recommendation system that generates responses to user reviews either automatically, or with some help from the developers. The proposed response recommendation system is a normative system that will generate responses that abide by the norms identified in this work, and will also monitor potential norm violations (if the responses were to be modified by the developers). Development of such a system forms the focus of future work.

Keywords: app reviews, norms, mining, developer responses, communication norms

1 Introduction

App reviews contain valuable information that can inform developers and users about issues that need to be fixed and enhancements and new features required [1]. The

Appears in the proceedings of COIN 2017 workshop held in conjunction with AAMAS 2017

availability of 'big-data' comprising users' reviews and developers' responses has provided an opportunity to study the type of communication norms (patterns of excepted social behavior in communication exchanges, also known as communication etiquettes [2, 3]) among involved parties (users and developers), to understand their prevalence and to make informed decisions about adopting them. While several studies have investigated the nature of app reviews submitted by users [1, 4, 7], only a few have investigated the nature of responses provided to these reviews by app development firms [5, 6]. Additionally, these works have not investigated response provision from a normative perspective. This work aims to bridge this gap by investigating the patterns of developers' responses to users' reviews.

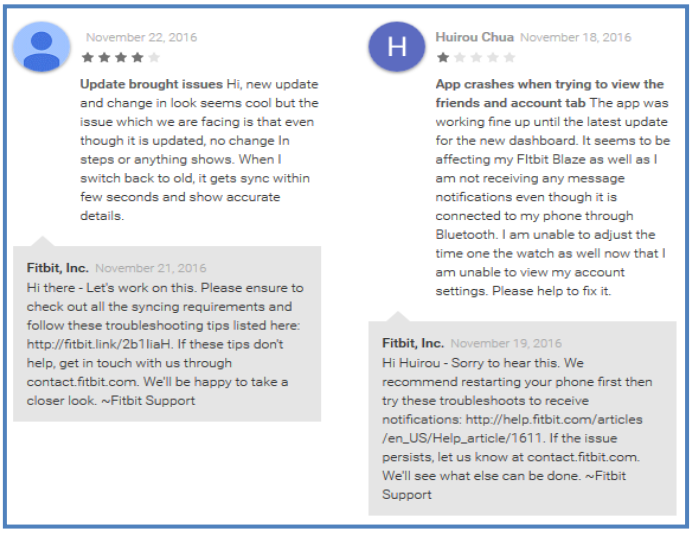

Fig. 1. Sample reviews from Fitbit app

Providing responses to reviews has only been a recent development, with Google Play enabling this service from 2013 [5, 6] (see examples of reviews and responses in Figure 1). This new functionality provides opportunities for researchers exploring the nature of responses provided, particularly from a norms perspective. It is important to study the presence of norms and their adoption levels in the new domain of app reviews for two reasons. First, insights on prevalent communication norms between users and developers can be inferred. In this work, we evaluate these aspects using two metrics: norm *awareness* and norm *adoption* (discussed in the next section). Second, the study helps to identify requirements for response recommendation systems, proposing responses to reviews. Such systems would solve the problem of developers having to respond to voluminous reviews [7]. Some apps attract tens of thousands of reviews each day (e.g., Pokemon Go received 40,000 reviews per day between its release and January 2017), and it is cumbersome and expensive for humans to respond to each of these reviews. One solution is to develop a response recommendation system that can generate appropriate responses based on historical data (i.e., re-

views and responses already available). The development of such a system needs the knowledge about prevailing communication norms. Once the system is aware of the prevailing communication norms it could respond appropriately considering such norms (i.e., a normative system).

This paper thus addresses three objectives: (1) it identifies a set of communication norms that can be identified in developer responses to user reviews, (2) it investigates norm awareness and adoption, and (3) it discusses the implications of the results for developing a normative response recommendation system. This paper is structured as follows. Section 2 provides background and related work considering communication norms. This section also introduces the domain of app development and how prior work has mined norms from large datasets including legal texts and business contracts. Section 3 presents the norms investigated in this study. Section 4 provides an overview of the adopted methodology, and Section 5 presents the results. Section 6 discusses the utility of the results in the development of a normative response recommendation system. Finally, Section 7 concludes the paper.

2 Background and related work

In communication studies, there has been a huge body of work focusing on the widely used email communication platforms [2, 3, 8-10] . From a norms perspective, researchers have focused on extracting patterns of common behaviors in email responses [3, 9, 10]. For example, the work of Kooti et al. [9] mined over 16 billion email interactions and noted that about 90% of emails are answered within a day. Research reported in [3, 10] focuses on different forms of greeting and closing norms in email messages. Beyond the email domain, researchers have focused on communication patterns in social-media platforms such as Twitter [11] and Facebook [12]. Kooti et al. [11] investigated the emergence and adoption of the retweet norm (i.e., the use of the 'RT' symbol for retweeting) when compared to the other proposals for retweeting. The work of Pérez-Sabater [12] investigated language conventions (greeting, closing, etc.) used by English and non-English speakers in Facebook posts.

In the context of software development, the domain of study of this work, there has been research investigating email discussions. For example, the work of Squire presents an overview of research work that uses email archives for various purposes including decision-making in software development [13]. Beyond emails, researchers have also investigated how users expressed their opinion on Twitter [14] about software they use (both desktop and mobile apps). Only recently, researchers have started investigating app reviews from a content-analysis viewpoint [1, 4]; we examine this body of work closely in the following sub-section.

2.1 User reviews and developer responses in app development

The app development domain has been popularized due to the rapid adoption of smart phones. After an app is released, users provide reviews that contain valuable information for other users and the developer(s) of the app. While prospective users read reviews to evaluate the suitability of apps for their needs, developers use these reviews for enhancing their app. The mechanism allowing users to provide reviews has been long implemented, however, the feature allowing developers to respond to reviews has only been enabled in app stores such as Google Play from 2013 (and this feature will be available for iOS only in the later part of 2017[1]). Since the provisioning of responses to reviews is relatively new, we believe norms in this domain might be in their formative stages, and hence, it is important to identify those norms so as to provide appropriate responses that conform to these norms.

Researchers have investigated the nature of reviews provided by users [1, 4, 7]. Maalej et al. [1] for instance have developed an approach that identifies whether reviews contain a bug report, a new feature request or a praise. The work of Chen et al. [4] proposes an approach for identifying informative reviews from noisy data and clustering reviews into relevant groups. Panichella et al. [7] have identified categories of users' requests (e.g., information giving, seeking) from reviews, thus, extending the work in [1]. However, developers' responses to users' reviews have not received a lot of attention. Developers' responses to users' reviews are important from a customer relationship management point of view since they: (a) indicate to the users that their feedback is listened to, appreciated, and appropriate action is taken to address their concerns, (b) provide solutions to problems, (c) help avoid losing customers [15] (reducing churn) which is wide-spread in the app community due to a large number of choices available, and (d) increase the reputation of brands and attract more customers (e.g., in the hotel industry, up to 60% increase in room sales as a result of providing responses (when compared to no responses) has been reported [16]). A few studies have investigated responses to app reviews [5, 6]. The work of McIlroy [5] noted that providing responses had a positive impact on user rating (about 38% attracting improved rating). The work of Bailey [6] investigated how developers accommodated user reviews (i.e., what actions developers undertook). However, the actions of developers weren't explicitly communicated to the users as the work investigated apps in iOS store which did not have the response feature for direct communication. Nevertheless, this study showed the usefulness of reviews for informing developers about their product deficiencies. We note these works have not scrutinized communication norms in users-developers interactions, as we do in this paper.

2.2 Mining norms in normative multi-agent systems

The perspective of norms that we adopt in this work has been borrowed from the normative multi-agent systems' research work [17], where norms are treated as prohi-

[1] https://techcrunch.com/2017/01/24/apple-will-finally-let-developers-respond-to-app-store-reviews/

bitions, obligations and permissions. Researchers have identified applications of these types of norms in many different domains [18]. For example, Gao and Singh have investigated a corpus of business contracts [19] that contain explicit specifications of prohibitions and obligations. Hashmi has proposed a methodology for extracting legal norms from regulatory documents [20]. Sadiq and Governatori [21] have discussed how various works have investigated norm compliance by investigating the extent to which business processes followed in organizations confirmed to norms (contractual obligations or legal requirements). Norms from textual documents in the abovementioned works (e.g., contracts [19]) are mined by the presence of explicit deontic modalities such as *must* and *must not* in sentences. For example, the phrase "One must pay his/her bills on time" would indicate the presence of an obligation. Researchers have also explored mining norms of software development from open source software repositories [22-24], applying and extending techniques previously developed for use in the multi-agent systems domain. For example, the work of Avery [22] scrutinizes the presence of different types of norms (prohibitions and obligations) in bug reports submitted by users and developers. Despite these developments, there is a gap in extracting and studying communication norms between users and developers in the software development context, a gap we address in this study. In the next section, we present the norms investigated in this work.

3 Norms investigated

We investigate three types of norms in this work: (1) *domain-specific response norms* (norms 1-4 in Table 1), (2) *obligation norms* (norms 5-11 in Table 1), and (3) *prohibition norms* (norm 12 in Table 1). A brief description of all the 12 norms investigated in this work is provided in Table 1.

Domain-specific response norms - The domain-specific response norms quantify the aggregate patterns in the domain of investigation (i.e., for all apps). These include *response rate* to reviews, *timeliness of responses*, *response length* and *review modification rate* (after the response is provided by the developer).

Obligation norms - These norms describe communication patterns that are expected to be followed. For example, a response is expected to contain a greeting [3]. We divide obligation norms into three parts – (1) the *social etiquette norms*, (2) *direct help norms*, and (3) *extended help norms*. These have been divided into three parts based on prior work on social etiquette norms in communications [2, 3, 9, 10], and content-analysis based works that have scrutinized the nature of responses provided to customers (direct help and extended help in other domains such as hotel reviews [15, 25, 26] and film reviews [27]). We describe these three aspects below.

1) Social etiquette norms correspond to social aspects of communication, including the use of greeting words such as 'Hi' and 'Dear' [3, 10], the use of customer name in responses (personalization) [3, 10], appreciating feedback provided [28], apologizing

for a complaint about the app [29], and sign-off (or closing) using developer's name or organization's name [3, 10]. These are norms 5-9 in Table 1. Figure 1 shows two sample responses. The response on the right conforms to the five norms mentioned above.

ID	Norm	Description (in the question form)
Category 1 - Domain specific response norms		
1	Response rate	What proportion of reviews received a response?
2	Timeliness of responses	How long (in days) does it take developers to respond to reviews?
3	Response length	What is the length of the reviews (in characters)?
4	Response modification rate	What proportions of reviews are modified after receiving a response?
Category 2 - Obligation norms		
(5-9)	***Social Etiquette norms***	
5	Salutation (opening)	Do responses contain appropriate salutation (e.g., Hi and Dear)?
6	Personalization	Do responses include user's name?
7	Appreciation	Do responses exhibit appreciation for feedback provided?
8	Apology	Do responses contain apologies for the inconveniences experienced by the users?
9	Personalized sign-off	Do responses have a personalized sign-off (e.g., developer name or organization name)?
10	***Direct help norms***	
a)	Provides complete or partial solution	Do responses provide solutions to the users' problems with the app?
b)	Indicates that solution will be available in the future or provides no solution	Do responses indicate that a solution will be available in the future or that a solution won't be available (for whatever reason)?
c)	Provides additional information/asks question/provides advise	Do responses contain additional information for the user, or offer advice or asks questions and provide answers?
11	***Extended help norms***	
a)	Details in a webpage (URL)	Does the review provide a URL of a web page for seeking additional help?
b)	Email	Does the review contain an email address for seeking additional help?
c)	Phone	Does the review contain a phone number for seeking additional help?
Category 3 - Prohibition norm		
12	Use of canned responses (i.e., same response to reviews)	Do developers use canned responses?

Table 1. Norms investigated in this work and their descriptions

2) Direct help norms correspond to the help offered towards solving an issue faced by the customer (norms 10a-c), or informing the customer that a solution will be availa-

ble in the future. A direct help norm addresses a problem in full or part (norm 10a), or provide information about whether the solution will be available in the future (norm 10b) and/or offers helpful suggestions for the user (norm 10c). These norms are inspired by the findings of researchers in other domains (e.g., hotel reviews [15, 25, 26] and film reviews [27]) who have studied the nature of reviews. The responses in Figure 1 provide direct help to the users.

3) *Extended help norms* (norms 11a-c) offer additional help to the customer. This help may be in conjunction with the direct help provided, or may be the only type of help offered. Literature on customer support channels notes there are numerous ways to listen to the customer such as phone, email, discussion boards and online chat. In fact, social-media platforms such as Twitter [30] also provide an opportunity to receive users' comments and facilitate service providers' responses. Content analysis of responses revealed the use of three channels for obtaining further help. These are: (1) a URL to a web page that can be used to communicate with the developers such as filling a web-based form (norm 11a), (2) an email address to write to the developers to obtain further help (norm 11b), and (3) a phone number to speak to a support person (norm 11c). The responses in Figure 1 contain URLs to the support web pages.

Prohibition norms – These proscribe certain actions or events. The prohibition norm that we investigate is the norm against using canned responses[2] that are perceived to be too impersonal (norm 12).

While some of the identified norms have been shown to hold in other relevant computer-mediated communication domains (e.g., social etiquette norms in email, and direct help norms in hotels' responses), we investigate whether these norms hold in the app review domain involving interactions between users and developers. We scrutinize this using two metrics – norm awareness and norm adoption. *Norm awareness* identifies whether the developers of an app are aware of a particular norm. If at least one of the responses from an app shows that the developer is aware of a norm, then the app is considered to be norm aware. *Norm adoption* is the extent to which a norm is adopted in the responses provided, taking into account the number of times the norm is adhered to in a situation where it is expected to hold. For example, if there are 20 responses provided by an app, and if four of these contain greeting words, then the app developers are assumed to be aware of the greeting norm (norm 1), and the norm adoption is 20% (4 out of 20).

4 Methodology

This study investigated norms in one particular domain of apps – the top-20 apps on Google Play that tracked running (i.e., running apps). The apps considered in this

[2] Examples of prohibitions of canned responses are listed in the guidelines for reviews by Amazon and TripAdvisor (refer to: https://www.amazon.com/gp/community-help/official-comment-guidelines and http://www.reviewtrackers.com/ultimate-guide-responding-tripadvisor-reviews/)

study are shown in Table 2. A Java program was implemented to obtain the reviews and responses for these apps as ranked by Google on 20th November 2016. We obtained a total of 24,408 reviews of which 2,668 had responses. The number of reviews and responses for each app are shown in columns 3 and 4 of Table 2. We analyzed the 2,668 responses to study the nature of response norms across the 20-apps. To identify response norms, we initially automated the identification process for norms 1-9 and 12 using customized SQL queries, utilizing keywords identified by the first author (in consultation with the second and third authors), based on previous research [2, 3, 8-10, 26], and also based on content analysis pursued on the dataset (use of hi, hey, dear, hello etc. for the greeting norms). While norms 1-9 and 12 could be fully automated using a keyword based approach, norms 10 and 11 required a manual interpretation due to the nature of the arguments presented in natural language.

In the process of manually interpreting the presence of norms 10 and 11, the presence of norms 5-9 were also validated. Note that the other norms (norms 1-4 and 12) do not require manual verification since the automation captures the metrics for these norms accurately (i.e., norms 1-4 report aggregated values and norm 12 checks for duplicate responses). To facilitate the verification process, we developed a tool which was used by two evaluators to indicate whether a review contains a particular norm (presence or absence of a norm). The first and the third author of this study divided the data into two parts to manually label the data for the presence of norms 5-11. For data analysis, we followed the interpretation guidelines for qualitative data [31]. The first author developed guidelines for coding the data, which was reviewed by the second author. After discussions between the two coders on how to interpret the presence of these norms, a sample of 20 developers' responses was coded and the results were discussed and adjustments were made to the interpretation guidelines. To formally evaluate the coding strength, a sample of 50 developers' responses was then coded by both investigators. The inter-rater reliability was computed using Cohen's kappa statistic measure which yielded a value of 0.94 for the two raters, which shows an almost perfect agreement (a value between 0.81 and 1 is considered to be almost perfect) [32]. The next section presents the results of our analysis of developers' responses.

5 Results

This section presents results for the three types of norms. The discussion of these results is presented in the subsequent section.

5.1 Domain specific response norms

The results for the domain specific response norms are given in Table 2.

Response rate (norm 1) - The overall response rate across the apps is 11%. This figure in closer to the previous finding of 13.8% by other researchers [5]. It can be observed that the response rate varies across the top-20 apps (min=0, max=62, S.D=22), as shown in column 5 of Table 2.

Timeliness (norm 2) – We observed that 81% of responses across all apps were provided within 7 days (see column 6 of Table 2) and 75% responses were provided within 4 days. However, only 21% of the responses are provided within the first day. This is in contrast to email communication norms where 90% of the responses were provided within the first day [9].

App ID	App Name	No. of reviews	No. of responses	Response rate	Timeliness (<7 days)	Response length	Review modification rate
1	Adidas train & run	805	252	31%	83%	293	12%
2	C25K® - 5K Running Trainer	1816	0	0%	0%	0	0%
3	Couch to 5K	732	4	1%	83%	118	0%
4	Endomondo - Running & Walking	1858	200	11%	84%	240	15.8%
5	FITAPP - Running Walking Fitness	196	121	62%	70%	151	0.6%
6	Fitbit	1161	339	29%	88%	278	12.4%
7	Fitso Running & Fitness App	389	83	21%	61%	240	11.4%
8	Garmin Connect Mobile	1263	150	12%	53%	217	39.8%
9	Google Fit - Fitness Tracking	1473	45	3%	82%	337	17.6%
10	Nike+ Run Club	1402	0	0%	0%	0	0%
11	Run With Map My Run	2079	0	0%	0%	0	0%
12	Runkeeper - GPS Track Run Walk	2006	8	0%	44%	204	55.6%
13	Running Distance Tracker	1150	709	62%	95%	55	2.3%
14	Running For Weight Loss	709	58	8%	90%	215	3%
15	Runtastic Running & Fitness	3146	28	1%	100%	132	0%
16	S Health	886	607	69%	89%	239	11.2%
17	Sportractive GPS Running App	94	10	11%	100%	122	10.5%
18	Sports Tracker Running Cycling	1274	13	1%	68%	143	0%
19	Strava Running and Cycling GPS	1727	0	0%	0%	0	0%
20	Zombies, Run! (Free)	241	41	17%	96%	185	4.4%
	Overall	**24407**	**2668**	**11%**	**81%**	**159**	**12%**

Table 1. Results of the domain specific response norms for the top-20 running apps

Response length (norm 3) – The average response length across the apps is 159 characters (min=55, max=337, S.D=105), which is less than half of the allowed response length of 350 characters. It can be observed from column 7 of Table 2 that some apps (app 1 and app 9) use the response length allowed more effectively (e.g., the reviews

contained detailed information) than others (with average response lengths of 293 and 337 respectively).
Review modification (norm 4) – After the responses are received users have modified 12% of the original reviews they had provided (min=0, max=56, S.D=15). This modification included changes to the reviews and the review rating. We noticed that the average rating for reviews that were modified was higher (average=3.12) than the ones that weren't modified (average=2.98), pointing towards the utility of the responses in satisfying users and enhancing their feelings about apps, ultimately resulting in apps attracting higher ratings.

5.2 Obligation norms

The results for the social etiquette norms, direct help norms and extended help norms (norms 5-11) are presented below.

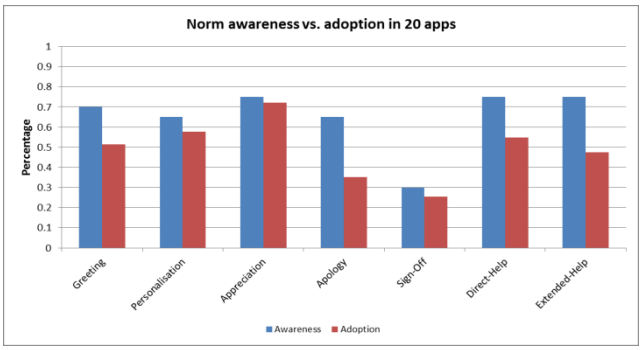

Fig. 2. Results for norm awareness vs. adoption in 20 apps

Social etiquette norms (norms 5-9) - We present the results based on two metrics – norm awareness and norm adoption. Results for norm awareness show that overall, developers of 70% of the apps are aware of the greeting norm (norm 5) and developers of 65% of the apps are aware of the personalization norm (norm 6). Eighty percent (80%) of the developers' responses towards reviews for the apps showed evidence for the awareness of the appreciation norm (norm 7). Awareness of apology norm was seen in 65% of the apps responses (norm 8). However, only 30% of the apps responses showed evidence for sign-off norm (norm 11). The results for norm adoption show that overall, 51% of responses had greetings (norm 5). The use of customers' name was present in 58% of responses (norm 6). In addition, 72% of responses contained appreciation (norm 7). However, only 35% of responses to reviews that had complaints contained apology (norm 8). Further, only 25% of the responses had a personalized sign-off (norm 11). The comparison for norm awareness versus adoption for the norms is shown in Figure 2 (including results for direct and extended help norms). The results for norms 5-11 show that while the apps developers are aware of these norms (average of 65%), their adoption rate is low (average of 49%).

Direct help norms - Figure 3 shows the results for the nature of direct help messages (norm 9) across apps for three categories – (1) solution available, (2) solution unavailable or will be available in the future, and (3) other helpful information. It can be observed that out of 16 apps that provided responses (excluding apps 2, 10, 11 and 19 that had zero responses), 13 of them (81%) had a higher proportion of responses that contained solution (category 1) than the other two categories (categories 2 and 3). Only in two apps (apps 5 and 7) there were more 'other' help information than the other two categories. App 9's responses did not contain any direct help to the user. These results show that responses in general have pointed to the solutions for issues faced by the users. It is interesting to note that some apps have provided help messages to the fullest extent possible – i.e., 100% (apps 12 and 18), although, they had responded only to a small number of reviews (8 and 13 respectively). Overall, norm awareness for direct help messages was 75%, and norm adoption was 55% (results shown in Figure 2).

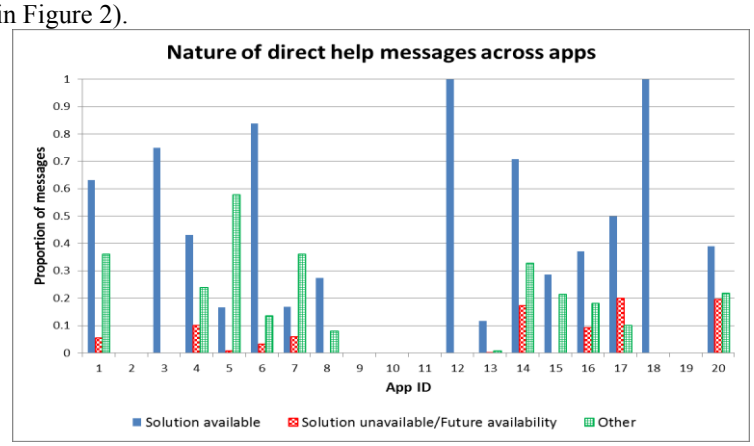

Fig. 3. Proportion of three types of direct help norms in responses across 20 apps.

Extended help norms - Figure 4 shows the results for the provision of extended help messages by developers across the apps (norm 11). Messages contained extended help information to contact support staff through a web interface, an email or phone reference. It can be observed that 15 out of 16 apps that provided responses (94%) used one of these three mechanisms for providing extended help, with 6 of them (38%) using two forms. Web interfaces and email addresses were used by 10 apps (63%), and the phone reference was least popular among the three options with only two apps using the option. However, the extent to which these extended help mechanisms were used varied. While apps 6 and 9 provided the URL for the users in more than 85% of the messages, apps 7 and 8 provided email addresses for a similar proportion of messages. These results show that the responses in general contained helpdesk details in an attempt to resolve the issues faced by users. Overall, norm awareness for providing extended help messages was 75% and norm adoption was 47% (shown in Figure 2).

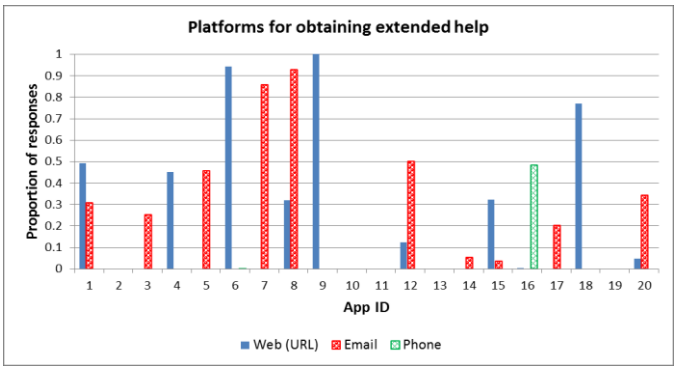

Fig. 4. Proportion of three types of extended help norms in responses across 20 apps.

5.3 Prohibition norms

Canned responses (norm 12) - There were about 30% (810 out of 2,668) of canned responses (reuse of the same responses). These messages did not have any personalization. Eight out of 16 apps used some form of canned responses (i.e., norm violated in 50% of the apps), with some apps using them excessively. For example, 75% or more of posts from apps 8 and 13 contained canned responses and 100% of app 9 responses (45 responses) were the same canned response.

6 Discussion

In this section we provide a discussion of the results presented in Section 5, particularly discussing the implications of the results presented in the previous section for developing a recommendation system that provides responses to reviews which conforms to norms.

a) Domain specific response norms – The results for the response rate norm (norm 1) show that only 11% of the reviews are responded to by developers. Prior research suggests that one of the key advantages of social media platforms for businesses is their ability to facilitate dialogue with their customers (e.g., keeping them informed about actions taken based on their feedback). The low response rate presents an opportunity to develop a response recommendation system that generates responses to new reviews, thus improving response rates. The recommendation system can directly post responses to reviews in straightforward cases and also provide a template of a response for a human to complete in complex cases. While some app developers have indeed realized the importance of responding to reviews (apps 5, 13 and 16) and have responded to more than 60% of reviews, certain other apps have completely ignored responding to reviews (apps 2, 10, 11 and 19). Research has shown that there is tendency for apps receiving a lot of reviews [5] to ignore responding to issues raised due to the voluminous nature of reviews. This again points towards the opportunity to develop a response recommendation system. The result of the timeliness norm (norm 2) shows that 96% of reviews that receive responses attract responses within 7 days. However, only 21% are responded to within a day (unlike 90% of emails responded to within a day [9]). The delay in providing response could be because of reasons such

as the time required to reproduce and then resolve a reported issue, and the availability of personnel. Having said that, research has shown that customer feedback needs to be responded to at the earliest since users' satisfaction and future repurchase intentions are directly related to the time taken to respond [33]. Additionally, in the app domain, users can easily try competitors' apps, hence, the timeliness of response needs improvement. A response recommendation system may be beneficial to provide responses in a timely fashion. The average response length for a message (norm 3) was 159 characters as opposed to 350 characters allowed. This also offers an opportunity to provide more detailed responses, which can be incorporated in generated responses (e.g., providing extended help as a part of every message and incorporating social etiquette norms). Our results for reviews' modification show that only 12% of reviews are modified after reading a response. Other researchers have noted that a significant proportion of users increased the ratings after reading a response, up to some 38% [5]. Hence, improving response provision rates through the proposed response generation system is likely to improve ratings.

b) Obligation norms – The results for the *social etiquette norms* (norms 5-9) indicate that norm adoption lags behind norm awareness. While app developers are aware of these norms, they do not adopt these norms. Independent of the reasons for the failure to adopt norms, norm adoption can be improved using a response recommendation system that considers the social etiquettes (greeting, use of customer names, thanking, apologizing, and signing-off). A recommendation system (or a software agent), can be easily programmed to adhere to these norms. *Direct help norms* (norms 10a-c) had the highest proportion of adoption (55%), among all norms. This is understandable given that the main goal of providing a response is to help users to overcome the issues they face. However, the proportion of responses that contain helpful information can be increased further. Two approaches may be beneficial towards this end. The first is to use appropriate machine learning approaches for identifying responses provided by the developers in the past for issues that are similar to the ones reported in the new review (i.e., review for which a response needs to be generated) similar to the work in [34], that recommends auto completion phrases for creating new reviews. If issues reported in a review match a previous review for which a response is available (e.g., based on a pre-determined threshold for matching features), the response generated can be directly posted to the user. The second approach involves the creation of template messages by developers for new issues that the users face (e.g., a solution describing a fix for a new bug that has not been reported by users before). A response recommendation system needs to consider both of these approaches (as presented in the basic workflow of the normative response recommendation system in Figure 5). *Extended help norms* (norms 10a-c) are easy to automate since they provide generic helpdesk information. However, these need to be customized based on the type of support (web, email or phone) the app development firm wishes to provide.

c) Prohibition norm – While the adoption of obligation norms can be ensured during the implementation of a response system, care should be taken to avoid violating prohibition norms. It is expected that developers will be presented with a template of a response which they can choose to modify. As a part of modifying the response, users may violate prohibitions. When these are violated, the system should warn the devel-

oper about potential issues. For example, if the developer replaces the response suggested by the system with an impersonalized canned response (norm 12), a warning can be generated. Note that 30% of the responses were canned responses. These responses do not use any personalization (i.e., without the use of greeting, customer name and developer sign-off). However, these elements can be easily incorporated in a response recommendation system (i.e., through the adherence of obligation norms 5-9). Additionally, warning messages can be generated if these norms are not adhered to when the developer posts the message. Thus, the response recommendation system should have a module that checks for violation of both types of norms (obligations and prohibitions) and provide appropriate warning messages to the developers.

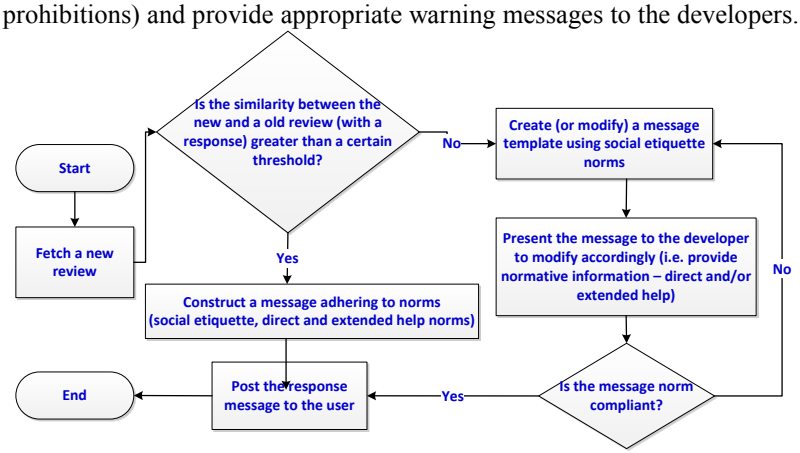

Fig. 5. Basic workflow of the normative response recommendation system

Further considerations on the normative response recommendation system – In addition to ensuring that the messages are norm compliant as shown in Figure 5 the system will also need to *prioritize* which reviews should be responded to first by the developers (e.g., based on urgency inferred through rating, and also whether other users face the same problem). For example if 10 users had rated the app as being poor and 8 of them complained about a new bug and two about a known bug (for which the template solution exists) then the system should present the response addressing the new bug issue to the developer to make a decision due to its potential impact on multiple users. The recommendation system should also be provided with some *autonomy* to post responses to certain messages without the intervention of the developers (e.g., all reviews that have 5 star ratings that do not have any suggestions for improvement or questions can be thanked automatically), considering the volume of reviews posted for the most popular apps. Additionally, if the responses generated do not meet the threshold for automatic post, they can be presented to the developers in the form of a ranked-ordered (priority) list for their action and approval, taking into account considerations such as urgency. If a similar response has to be provided to multiple users, rephrasing responses can be considered using existing services[3] to avoid the issue of canned responses. Although, personalization of messages using appropriate

[3] http://www.gingersoftware.com/

usernames and sign-offs is likely to somewhat reduce the canned response problem. Implementing the outlined solution forms the focus of our subsequent work.

7 Conclusion

In the domain of app development, communication norms between users and developers have rarely been investigated. This work addresses this gap by first abstracting three categories of response norms for app reviews comprising 12 norms (4 domain specific response norms, 7 obligation norms and 1 prohibition norm). It scrutinizes awareness and adoption of these norms from a dataset of user reviews and developer responses in Google Play's top-20 running apps. The work demonstrated that there is a gap between norm awareness and adoption across apps. App developers, despite being aware of these norms, do not adopt the norms effectively. To improve the provision of responses to users this work proposes a normative response recommendation system that also has the potential to improve user satisfaction, increase user ratings and potentially reduce developers' workload in responding to reviews.

8 References

1. Maalej, W., Nabil, H.: Bug report, feature request, or simply praise? on automatically classifying app reviews. In RE, 2015 conference, pp. 116-125. (2015)
2. Lewin-Jones, J., Mason, V.: Understanding style, language and etiquette in email communication in higher education: a survey. Research in PC Education 19, 75-90 (2014)
3. Waldvogel, J.: Greetings and closings in workplace email. Journal of Computer-Mediated Communication 12, 456-477 (2007)
4. Chen, N., Lin, J., Hoi, S.C., Xiao, X., Zhang, B.: AR-miner: mining informative reviews for developers from mobile app marketplace. In ICSE conference, pp. 767-778. ACM, (2014)
5. McIlroy, S., Shang, W., Ali, N., Hassan, A.: Is it worth responding to reviews? a case study of the top free apps in the Google Play store. IEEE Software (2015)
6. Bailey, K.: Out of the Mouths of Users: Examining User-Developer Feedback Loops Facilitated by App Stores. (2015)
7. Panichella, S., Di Sorbo, A., Guzman, E., Visaggio, C.A., Canfora, G., Gall, H.C.: How can I improve my app? Classifying user reviews for software maintenance and evolution. In ICSME conference, pp. 281-290 (2015)
8. Markus, M.L.: Finding a happy medium: Explaining the negative effects of electronic communication on social life at work. ACM TOIS 12, 119-149 (1994)
9. Kooti, F., Aiello, L.M., Grbovic, M., Lerman, K., Mantrach, A.: Evolution of conversations in the age of email overload. In WWW conference, pp. 603-613 (2015)
10. Bjørge, A.K.: Power distance in English lingua franca email communication1. International Journal of Applied Linguistics 17, 60-80 (2007)
11. Kooti, F., Yang, H., Cha, M., Gummadi, P.K., Mason, W.A.: The Emergence of Conventions in Online Social Networks. In ICWSM conference (2012)
12. Pérez-Sabater, C.: The linguistics of social networking: A study of writing conventions on facebook. Linguistik online 56, (2013)
13. Squire, M.: How the FLOSS research community uses email archives. International Journal of Open Source Software and Processes (IJOSSP) 4, 37-59 (2012)

14. Guzman, E., Alkadhi, R., Seyff, N.: A Needle in a Haystack: What Do Twitter Users Say about Software? In RE conference, pp. 96-105 (2016)
15. Chan, N.L., Guillet, B.D.: Investigation of social media marketing: how does the hotel industry in Hong Kong perform in marketing on social media websites? Journal of Travel & Tourism Marketing 28, 345-368 (2011)
16. Ye, Q., Gu, B., Chen, W., Law, R.: Measuring the value of managerial responses to online reviews-a natural experiment of two online travel agencies. In ICIS conference (2008)
17. Andrighetto, G., Governatori, G., Noriega, P., van der Torre, L.W.: Normative multi-agent systems. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2013)
18. Singh, M.P., Arrott, M., Balke, T., Chopra, A.K., Christiaanse, R., Cranefield, S., Dignum, F., Eynard, D., Farcas, E., Fornara, N.: The uses of norms. Schloss Dagstuhl LZI (2013)
19. Gao, X., Singh, M.P.: Extracting normative relationships from business contracts. In AAMAS conference, pp. 101-108. (2014)
20. Hashmi, M.: A methodology for extracting legal norms from regulatory documents. In EDOCW workshop, pp. 41-50 (2015)
21. Sadiq, S., Governatori, G.: Managing regulatory compliance in business processes. Handbook on Business Process Management 2, pp. 159-175. Springer (2010)
22. Avery, D., Dam, H.K., Savarimuthu, B.T.R., Ghose, A.: Externalization of software behavior by the mining of norms. In MSR conference, pp. 223-234 (2016)
23. Dam, H.K., Savarimuthu, B.T.R., Avery, D., Ghose, A.: Mining software repositories for social norms. In ICSE conference, pp. 627-630 (2016)
24. Savarimuthu, B.T.R., Dam, H.K.: Towards mining norms in open source software repositories. In ADMI workshop, pp. 26-39 (2013)
25. Ye, Q., Law, R., Gu, B.: The impact of online user reviews on hotel room sales. International Journal of Hospitality Management 28, 180-182 (2009)
26. Sparks, B.A., So, K.K.F., Bradley, G.L.: Responding to negative online reviews: The effects of hotel responses on customer inferences of trust and concern. Tourism Management 53, 74-85 (2016)
27. Basuroy, S., Chatterjee, S., Ravid, S.A.: How critical are critical reviews? The box office effects of film critics, star power, and budgets. Journal of marketing 67, 103-117 (2003)
28. Khoo-Lattimore, C., Ekiz, E.H.: Power in praise: Exploring online compliments on luxury hotels in Malaysia. Tourism and Hospitality Research 14, 152-159 (2014)
29. Matzat, U., Snijders, C.: Rebuilding Trust in Online Shops on Consumer Review Sites: Sellers' Responses to User-Generated Complaints. Journal of CMC 18, 62-79 (2012)
30. Malthouse, E.C., Haenlein, M., Skiera, B., Wege, E., Zhang, M.: Managing customer relationships in the social media era: introducing the social CRM house. Journal of Interactive Marketing 27, 270-280 (2013)
31. Merriam, S.B., Tisdell, E.J.: Qualitative research: A guide to design and implementation. John Wiley & Sons (2015)
32. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics 159-174 (1977)
33. Miller, R.B.: Response time in man-computer conversational transactions. In FJCC conference, pp. 267-277 (1968)
34. Arnold, K.C., Gajos, K.Z., Kalai, A.T.: On Suggesting Phrases vs. Predicting Words for Mobile Text Composition. In UIST symposium, pp. 603-608 (2016)