

Now, Later, Never: A Study of Urgency in Mobile Push-notifications

Beatriz Esteves¹[0000-0003-0259-7560], Kieran Fraser²[0000-0002-5363-0706],
Shridhar Kulkarni²[0000-0001-8077-4011], Owen Conlan²[0000-0002-9054-9747],
and Víctor Rodríguez-Doncel¹[0000-0003-1076-2511]

¹ Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain
`beatriz.gesteves@upm.es`

² ADAPT Centre, Trinity College Dublin, Dublin, Ireland

Abstract. Push-notifications, by design, attempt to grab the attention of subscribers and impart new or valuable information in a particular context. These nudges are commonly initiated by marketing teams and subsequent delivery interruptions tend to conflict with subscriber priorities and activities. In this work, we present a definition of urgency applied to notifications. We describe its value in an ontology for push-notification annotation and also evaluate a variety of classification models tasked with distinguishing urgency levels in notification text. The best model achieved an F1-score of 0.89. The proposed models have the potential to benefit subscribers by helping them better prioritize incoming notifications and also aid marketers in creating time-relevant campaigns.

Keywords: Push-notification · Urgency · Semantic Web · Multi-label Text Classification · Marketing

This preprint has not undergone any post-submission improvements or corrections. The work was accepted and is going to be presented at 20th International Conference on Advances in Mobile Computing Multimedia Intelligence, MoMM 2022, Virtual Conference, November 28-30, 2022. The Version of Record of this contribution will be published in the Advances in Mobile Computing and Multimedia Intelligence volume (MoMM 2022, LNCS 13634 proceedings), and will be available online at https://link.springer.com/chapter/10.1007/978-3-031-20436-4_4.

1 Introduction

Push-notifications were first used as a mechanism for alerting email users they had received a new message [12], with the intention of saving users' time and effort repeatedly checking for new emails. Almost 20 years later, non-urgent notifications are still pushed and delivered at the discretion of apps and marketing teams, with little regard for subscribers. The situation today is much more difficult to manage as notifications are spawned from sources beyond the original desktop email client to include mobile devices and other IoT devices.

Research in the area of intelligent Notification Management has explored the relationship between notifications and user attendance in order to help subscribers prioritize their time with respect to incoming nudges. These methods for improving notification delivery depend on clear notification features which

express the intent and value of a notification in a given moment. Few notification features exist to explicitly express the time-sensitivity or the period of time for which a notification remains relevant which could be used by the subscriber to better prioritize notification attendance and could also be leveraged by the publishing app to update or remove notifications which have expired and may contain inaccuracies or misleading information due to the time lapse between delivery and action. In this work we define a novel feature which expresses the urgency and period of relevancy for mobile notifications and evaluate a range of classification models on their ability to predict the urgency of a notification using minimal input features.

2 Background

In their work examining the prioritization of emails by subscribers, Cox et al. [5] discuss the importance of urgency as a feature indicative of email response. The authors also highlight how urgency is but one contributing factor for deciding how to prioritize incoming emails, others including the time to respond, message importance and text ambiguity. This aligns with research by Zhu et al. [16] who proposed the “mere urgency effect”³.

Fraser et al. [7] studied the impact of this “information-gap” within push-notifications and emphasized the potential negative impact it could have if consistently used as a dark pattern by marketers for enticing engagement.

Quantifying urgency within communication channels has been explored by Kalman et al. [8] in their evaluation of “chronemic urgency of digital communication media”. The authors were able to identify specific traits for media channels associated with spawning messages of high chronemic urgency - which is the urgency perceived by the message receiver. Texting and calling channels were found to have the highest associated urgency and response time. Mehrotra et al. [11] and Acosta et al. [1] opted for a binary indicator to select whether the content of a notification was urgent or not and Aranda et al. [3] used a taxonomy of four different types of notifications that also considers their enjoyability.

Whilst research to date has attempted to indicate a level of urgency within messages pushed at subscribers, few have yet attempted to extract it autonomously from push-notification text and, currently, none include predictions for the period of time which the content remains relevant. This work attempts to bridge this identified gap by defining a standard set of urgency labels for push-notifications and evaluating a range of multi-label classification models on an urgency prediction task. In this context, the usage of a Semantic Web⁴ ontology seems perfectly aligned with the goals of this work, as it provides the foundation for it to be easily reused and extended by other researchers and connected with other ontologies. Although there are already some ontologies that model temporal aspects⁵, no works were found to model the time urgency of text pieces.

³ “people will be more likely to perform an unimportant task over an important task (...) when the unimportant task is merely characterized by *spurious* urgency”

⁴ For a complete definition see <https://www.w3.org/standards/semanticweb/>.

⁵ Time ontology: <https://www.w3.org/TR/owl-time/>

3 Methodology

3.1 Mobile Push-notification Dataset

A push-notification social listening tool, developed by EmPushy⁶, was used to collect a variety of features from notifications pushed in real-time over a period of 550 days. The social listening tool was subscribed to apps sourced from 37 categories of the Google Play Store providing a wide net to be cast for notifications associated with differing types and levels of urgency and relevancy. The social listening tool was deployed within the geographical region of Ireland and was run 24/7 to ensure maximal capture of pushed notifications. Whilst 673 apps were subscribed to during this period, only 525 (78%) were found to push notifications, indicating that a large portion of apps did not exploit the use of push-notifications for driving engagement and communicating with their subscribers. A visual representation of the components of a common mobile push-notification is available at <https://empushy.github.io/momm22/#push>.

3.2 Annotating Urgency with APN

As discussed in Section 2, at the time of writing, no standard feature exists which describe the urgency of a push-notification paired with an assumed period of relevancy. This work proposes a taxonomy to address this gap – APN⁷, first introduced to categorize the call-to-action of notifications [6], classifies urgency through a taxonomy of eight categories. Using such a vocabulary provides the opportunity to improve the transparency and explainability of models developed with the identified urgency labels, beyond the other advantages already mentioned in Section 2. A diagram with the different urgency categories defined by APN and their definitions can be found at <https://w3id.org/apn/#x4-urgency>. Moreover, examples of APN-annotated push-notifications are available in the ontology documentation at <https://w3id.org/apn/#x9-2-urgency-cta>.

3.3 Crowdsourced Annotation

As described previously, the notifications were collected using EmPushy’s social listening tool. As some apps tended to push more notifications than others, a balancing script was created to ensure that notifications selected for annotation were evenly distributed amongst app categories and individual apps within those app categories. In addition, to ensure a high level of diversity, the text of the notification content was combined and converted to a sentence embedding [13] then ranked using cosine similarity to include only the least similar notifications in the final data set for annotation.

For annotation, the Appen Platform⁸ was used as it provided a global workforce and self-service tool set for creating and managing the annotation task at scale. More information on the implemented process for annotation, including an illustration, is available at <https://empushy.github.io/momm22/#annotation>.

⁶ <https://www.empushy.com>

⁷ Online documentation for APN is available at <https://w3id.org/apn/>.

⁸ Crowdsourced annotation platform: <https://appen.com/>

4 Understanding Urgency in Mobile Push-notifications

525 applications pushed notifications during the 550 day period EmPushy’s social-listening tool was running. The subsequent 120,990 notifications logged were collected across 36 unique app categories⁹. A number of text features were extracted from the notifications to better model their text content. Python packages were used to engineer a number of text features which were shown to be statistically significantly different across varying app category types¹⁰. Naturally, from this we can conclude that marketers crafting the text content for campaigns do so differently depending on the type of app and audience they are targeting. This is important, as it suggests urgency could be represented in different ways within notification text content across differing domains and as such, should be included in any input to algorithms seeking to extract urgency autonomously.

5 Evaluating Urgency Classification

This paper thus far has discussed labeling push-notifications with associated urgency using a human crowd of annotators armed with a novel push-notification taxonomy. Whilst human-in-the-loop inference is helpful for understanding the relationships between context, notification features and urgency labels, it is not feasible for *every pushed notification* to be manually classified. Machine Learning (ML) has the potential to facilitate autonomous categorization to an associated urgency category and provide enhanced clarity for those creating campaigns and transparency for the subscribers.

5.1 Experiment 1 - Baseline

Related research has shown that ML has worked well at predicting the likelihood of a subscriber replying to an email [14], a notification being accepted based on its text content features [10] or at extracting the urgency level of emails to help prioritize work-place tasks [2]. In this work, we considered four algorithms for the task of classifying urgency in notifications: Naive Bayes, Random Forest, AdaBoost and XGBoost¹¹. The problem was framed as a multi-label classification task as the labels are not mutually exclusive, e.g., a notification may be categorized as both relevant for “weeks” and “season duration”.

The annotated data was split into train (80%) and test (20%) sets and two problem transformation approaches were applied for facilitating multi-label classification, the **Classifier Chains** and the **Binary Relevance** algorithms. The final output was the union of predictions made by each individual classifier. As a baseline experiment, only notification text and type features were used.

The performance of each algorithm is illustrated in the Figure available at <https://empushy.github.io/momm22/#baseline>. Even though the classifier-chain performed better than binary relevance in the AdaBoost and XGBoost cases, overall there was little discrepancy between the two approaches. In all

⁹ App push statistics available at <https://empushy.github.io/momm22/#app-stats>.

¹⁰ More information available at <https://empushy.github.io/momm22/#test-stats>.

¹¹ References available at <https://empushy.github.io/momm22/#algorithms>.

cases, the F1-score did not surpass 0.35, which indicated relatively poor classification performance, albeit on a new and challenging task. The results of this experiment however did provide a baseline on which to evaluate future approaches.

5.2 Experiment 2 - Data Augmentation

One hypothesis for the poor results of experiment 1 was that the quantity of data was insufficient for the task at hand. Shorten et al. [15] suggested that “synonym swapping” could strengthen the decision boundary and enable a classification model to better distinguish between target labels. Therefore, this approach was implemented to augment the text within the annotated notification data set with additional instances varying only by a few number of key words “swapped out” for synonyms. In total, 29,938 additional instances were created. In addition, a neural augmentation technique was also used to further increase the quantity of data available for training. Beddiar et al. [4] paired this technique with paraphrasing and found it helped to expand the quantity of data to 20 times the original size. In total, 13,124 additional instances were created using this technique.

Overall, applying data augmentation increased the quantity of annotated notification data by a factor of ≈ 4.28 , from 13,124 to 56,186 instances. Once the data augmentation step was completed, the same experimentation process was executed to train and evaluate the four models presented in the previous section. The performance of each algorithm is illustrated at <https://empushy.github.io/momm22/#augmentation>. The increased performance is stark compared to the baseline – the F1-score of all models increased significantly, with Random Forest being most improved from an initial score of 0.3 to 0.7.

5.3 Experiment 3 - Time Expressions

To further improve the performance of the classifier, it was hypothesized that time-related information contained within the notification could be extracted and used as an additional feature. There have been numerous research challenge tasks addressing temporal processing over the past number of years. Utilizing this body of work, time expressions were extracted from the annotated notification text and added as an additional feature to our models.¹²

Of the 13,124 notifications annotated with urgency labels, 5,020 (38%) were found to contain time-expression features (as defined in the *SCATE* schema [9]). Of the 63 time-expression labels, just 24 appeared in notification text. A table with the 10 most frequent time-expressions and a figure illustrating the performance of each algorithm are available at <https://empushy.github.io/momm22/#time>. Once again, all algorithms improved in performance. The F1-score of XGBoost in particular saw a $\approx 14\%$ improvement (from 0.69 to 0.79) indicating that time-expression data is indicative of notification urgency and provides utility to algorithms tasked with predicting it using notification text content.

¹² See a complete list of used works at <https://empushy.github.io/momm22/#time>.

5.4 Experiment 4 - Delivery Date

Due to the success of extracting time-expression information and its subsequent positive impact on urgency prediction, the final experiment hypothesized that notification delivery time would also improve the algorithms ability at predicting the urgency and relevance of a notification.

The performance of each algorithm, with the addition of the delivery time as an input feature, is illustrated at <https://empushy.github.io/momm22/#delivery>. Once again, all algorithms show improved performance with the additional information made available. There is little difference between problem transformation approaches, but binary relevance slightly improves performance in two cases. XGBoost once again was the top performing algorithm achieving a final F1-score of 0.89 (a $\approx 12\%$ increase over experiment 3).

6 Conclusions

This work explored the importance of timeliness in the delivery of notifications for both subscribers to receive relevant content on time and for marketers to create time-relevant campaigns. No research had been done so far to create a clear categorization for these services to use. APN's urgency taxonomy fills this gap by providing terms to specify the urgency and relevancy period of notifications.

In addition, four experiments were performed and evaluated to build a set of classification models capable of predicting notifications' urgency. XGBoost was the top performing algorithm, with a final F1-score of 0.89. As future work, different modes of notification services should be studied and evaluation based on direct user feedback should be performed.

Acknowledgements This research has been supported by the EU's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 813497 as well as with the financial support of Enterprise Ireland, the ERDF under Ireland's ESIF Programme 2014-2020 and SFI under Grant Agreement No 13/RC/2106_P2 at the ADAPT Centre at Trinity College.

References

1. Acosta, M.E., Palaoag, T.D.: Characterization of Disaster Related Tweets According to Its Urgency: A Pattern Recognition. In: ICCAI'19. pp. 30–37 (2019)
2. Alshehri, Y.A.: Tasks' classification based on their urgency and importance. In: ICCDA'20. pp. 183–189 (2020)
3. Aranda, J., Ali-Hasan, N., Baig, S.: I'm just trying to survive. In: MobileHCI'16. pp. 564–574 (2016)
4. Beddiar, D.R., Jahan, M.S., Oussalah, M.: Data expansion using back translation and paraphrasing for hate speech detection. OSNEM **24**, 100153 (2021)
5. Cox, A., Bird, J., Brumby, D., Cecchinato, M., Gould, S.: Prioritizing unread e-mails: people send urgent responses before important or short ones. HCI (2020)
6. Esteves, B., Fraser, K., Kulkarni, S., Conlan, O., Rodríguez-Doncel, V.: Extracting and understanding ctas of push-notifications. In: NLDB'22. pp. 147–159 (2022)
7. Fraser, K., Conlan, O.: Enticing Notification Text & the Impact on Engagement. p. 444–449. UbiComp-ISWC '20 (2020)

8. Kalman, Y.M., Ballard, D.I., Aguilar, A.M.: Chronemic urgency in everyday digital communication. *Time & Society* **30**(2), 153–175 (2021)
9. Laparra, E., Xu, D., Bethard, S.: From characters to time intervals: New paradigms for evaluation and neural parsing of time normalization. *TACL* **6**, 343–356 (2018)
10. Mehrotra, A., Musolesi, M., Hendley, R., Pejovic, V.: Designing Content-Driven Intelligent Notification Mechanisms. In: *UbiComp’15*. p. 813–824 (2015)
11. Mehrotra, A., Pejovic, V., Vermeulen, J., Hendley, R., Musolesi, M.: My phone and me. In: *CHI’16*. pp. 1021–1032 (2016)
12. Middleton, C.A., Cukier, W.: Is mobile email functional or dysfunctional? Two perspectives on mobile email usage. *EJIS* **15**(3), 252–260 (2006)
13. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: *EMNLP’19* (2019)
14. Sappelli, M., Verberne, S., Kraaij, W.: Combining textual and non-textual features for e-mail importance estimation (2013)
15. Shorten, C., Khoshgoftaar, T.M., Furht, B.: Text data augmentation for deep learning. *Journal of Big Data* **8**(1), 1–34 (2021)
16. Zhu, M., Yang, Y., Hsee, C.K.: The mere urgency effect. *Journal of Consumer Research* **45**(3), 673–690 (2018)