False Positive Intent Detection Framework for Chatbot Annotation

Lecia Lim
Analytics Center of Excellence, DBS Bank, Singapore

Samarth Agarwal
Analytics Center of Excellence, DBS Bank, Singapore

Xuejie Zhang
Analytics Center of Excellence, DBS Bank, Singapore

John Jianan Lu
Analytics Center of Excellence, DBS Bank, Singapore

ABSTRACT

For chatbots answering thousands of user queries daily, it requires huge annotation efforts or explicit signals from users to identify incorrect chatbot predictions. Identification of such False Positives is key to improving chatbot accuracy and is a challenging problem due to the high cost and limited explicit signals from users. In this paper, we present a framework for automatically detecting False Positive intents in an employee chatbot through implicit feedback by capturing specific user behavior using techniques such as detection of repeated queries and leveraging on active learning sampling strategies to find cases where the chatbot might have provided an incorrect response. Using this approach within the bank, annotators can prioritize their efforts and detect False Positive intent approximately three times better than manual screening of random chatbot dialogues. This framework can be reused across different chatbot applications.

CCS CONCEPTS

• **Information systems** → Information retrieval; Information retrieval query processing; Query log analysis.

KEYWORDS

chatbot, digital transformation, banking, false positive, duplicate detection, natural language processing, information retrieval, implicit feedback, active learning query strategies

ACM Reference Format:
Lecia Lim, Samarth Agarwal, Xuejie Zhang, and John Jianan Lu. 2022. False Positive Intent Detection Framework for Chatbot Annotation. In *2022 6th International Conference on Natural Language Processing and Information Retrieval (NLPIR 2022), December 16–18, 2022, Bangkok, Thailand*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3582768.3582798

1 INTRODUCTION

Many companies are strengthening the digital culture among their workforce and thoroughly modernizing the work environment through the "integration of digital technology into all areas of business, fundamentally changing how one operates and delivers value to customers." [1]. The use of chatbots is becoming ubiquitous, and they can be used as a tool to drive digital transformation by transforming the employee work experience. Chatbots can help to transform different business functions and are used in various use cases, such as answering questions in customer service [2], promoting products in marketing [3] and scheduling appointments in healthcare [4]. They are also a time-saving resource that can be used by employees to improve work productivity through automation and workload reduction [5]. Internal employee chatbots have the potential to boost accessibility, efficiency, and employee satisfaction in the workplace. It is especially useful in the self-service or support domain, where they carry out form-filling processes and supply information. Some examples of internal chatbot usage are answering questions (e.g., how to reset a password), providing information (e.g., employee onboarding), and simple day-to-day tasks such as requesting time off. While the chatbot can provide a multitude of benefits, the employee would not be able to enjoy these benefits if the intent prediction is inaccurate. Although there have been many research efforts on improving chatbot conversational capabilities [6–8], little has been done on the chatbot dialogue analysis and improvement process. Chatbot dialogue analysis is important as it allows one to improve on the chatbot's content and capability for intent recognition so that it may respond correctly to similar requests in the future [9]. Kvale et al. presented a comprehensive study of chatbot dialogue analysis and identified eight improvement areas to improve chatbot performance [9]. One improvement area is False Positive (FP), whereby the chatbot erroneously interprets the user query and predicts the wrong intent, and a suggestion is to update the chatbot training data.

2 BACKGROUND

The employee chatbot within the bank uses a Convolutional Neural Network (CNN) classification model trained from scratch on a set of questions and intents provided by various domains such as Human Resources (HR) and Information Technology (IT). Each domain has its own set of training data and intent class, and there are checks in place to ensure that multiple intents are not mapped to the same question. Text processing such as case handling, punctuation and stopword removal, spell correction, tokenization, and lemmatization is applied to the questions before passing into the model. When an employee asks a question, the chatbot returns a response that is mapped to an intent with the highest confidence score. If that score is less than 0.4, the query would be tagged as 'unknown', and the chatbot would express uncertainty or suggest multiple responses to enhance the capabilities of conversational repair in the chatbot [10]. Typically, if a chatbot is not confident in its intent prediction, it either responds with an answer that is

NLPIR 2022, December 16–18, 2022, Bangkok, Thailand

ACM ISBN 978-1-4503-9762-9/22/12.
https://doi.org/10.1145/3582768.3582798

incorrect or it returns a message stating that it does not understand the question. Both responses are frustrating to users as they are not sensitive to how conversational repair is carried out in human-to-human dialogues [11]. A side benefit of conversational repair is that one of the suggested responses might be the correct one, thus increasing the chatbot's accuracy and reducing its False Positives. False Positives can occur for a variety of reasons, including a lack of data for specific intents, overlap between multiple intents, or a model that is over-fitted to specific signals in the data. If the chatbot is unable to handle the enquiries adequately, it could result in a bad user experience and reduce chatbot usage in the long run [2, 12]. Therefore, it is imperative to identify and annotate the False Positives to enhance the chatbot training data and improve its intent prediction.

False Positives can be identified through explicit or implicit feedback. Explicit feedback is obtained when the user performs a specific action designed to give feedback to the system. For instance, a 'like' and 'dislike' button could be embedded within each response, and if the user clicks on 'dislike', the intent will be labeled as 'False Positive'. While the explicit feedback can provide a direct signal of the False Positive labels, users can choose to not participate in the feedback process if the feedback functionality is not designed as a compulsory field. As very little explicit feedback is collected in our use case, the existing approach to identify the potential False Positive intents is to use the least confidence query strategy in active learning. Queries of the low confidence responses (i.e., maximum confidence of intent prediction less than 0.4) are then annotated using a semantic search based smart annotation solution introduced by Agarwal et al. [13]. The annotated query-intent pair data is then added back into the chatbot training data for model retraining to improve the chatbot accuracy.

While False Positives might also exist among the high confidence cases, it is impossible for annotators to review all these cases due to the sheer number of records (the ratio of high:low confidence records are 24:1). Hence, the existing approach for identifying high-confidence False Positives is through random sampling. This process is inefficient, and the False Positive detection precision is very low. Annotators from the Business Unit (BU) reviewed 940 high-confidence records randomly sampled across a 2-week period, and only 224 were identified as False Positives. This translates to a baseline False Positive detection precision of 23.8%.

With limited explicit feedback, a huge challenge in identifying False Positive intents in chatbot dialogue data is that human annotators are required to manually screen through the dataset. To prioritize the annotators' efforts and improve the process of identifying False Positives, a framework was developed to automatically detect False Positive intents using implicit feedback by capturing specific user behavior using techniques such as duplicate detection. Additionally, we included more active learning sampling strategy (margin sampling and entropy sampling) to identify scenarios where the model is not confident in its intent prediction. From the online and offline testing, it can be concluded that this framework is able to attain a higher degree of False Positive detection precision for high confidence responses as compared to the existing approach of random sampling.

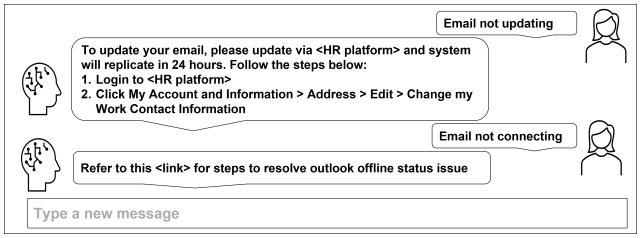

Figure 1: Example of a FAQ chatbot dialogue in a session

3 DATA

The chatbot transcript data is a record of the chat between the employee and the chatbot. Each row of record consists of the session id, conversation id, timestamp, conversation input, matched intent, intent confidence score, and remarks. The 'session id' and 'conversation id' columns are the session and conversation unique identifiers. The 'timestamp' is the date and time of the user's query. The 'conversation input' column contains the user's query or action. The 'matched intent' column includes the chatbot's intent prediction with the maximum confidence score. The 'confidence score' column is the corresponding maximum confidence score. The 'remarks' column contains a list of dictionaries of all intents and their respective confidence scores. Figure 1 and Table 1 show an example of the chatbot dialogue and its corresponding chatbot transcript data.

The 'matched intent' is tagged to a particular response, which could either be a Guided Conversation (GC) or a Frequently Asked Questions (FAQ). GC transforms traditional form-filling services into guided digital service completion by allowing employees to complete services with a few clicks. Figure 1 and Figure 2 shows an example of the chatbot dialogue with a FAQ and GC response, respectively. Table 1 and Table 2 shows the corresponding chatbot transcript log data. When annotators review and annotate the data, two additional columns are added: 'need_annotation' and 'correct_intent'. The 'need_annotation' column is marked as 1 when the record has a False Positive intent prediction, and the 'correct_intent' includes the true intent prediction. The annotated query-intent pair is passed back into the chatbot training data for model retraining.

4 FALSE POSITIVE DETECTION

To capture the implicit feedback, the user journeys (in Figure 3) were hypothesized to reflect the possible user behavior when the chatbot suggests an inaccurate response. For instance, a user may ask the same query in a different way, indicating dissatisfaction, or a user might exit the session altogether. The ideal flow of a GC response would require the user to complete the GC flow. Hence, if a GC response is not triggered and the user exits, it might signify dissatisfaction. However, in the case of a FAQ response, it is difficult to say if the user exited because he was discontent or if he got the required response. Apart from capturing implicit feedback, another indication of chatbot accuracy can be determined by the intent probabilities. If the chatbot is confident in its response, the intent predictions would be differentiable across intents and the maximum intent probability would be high. Conversely, multiple intents with a similar high confidence score could imply that the

Table 1: Chatbot dialogue transcript log data for Figure 1

session id	convo id	time stamp	convo input	matched intent	score	remarks
S123	C001	06-05-2022 8:30:32	Email not updating	hr.update_contact_info	0.621	[{intent: hr.update_contact_information, confidence:0.621, domain:HR}, {intent: ts.outlook_connectivity, confidence:0.555, domain:TS}, . . .]
S123	C002	06-05-2022 8:31:53	Email not connecting	ts.outlook_connectivity	0.785	[{intent: hr.update_contact_information, confidence:0.403, domain: HR}, {intent: ts.outlook_connectivity, confidence:0.785, domain:TS}, . . .]

Table 2: Chatbot dialogue transcript log data for Figure 2

session id	convo id	time stamp	convo input	matched intent	score	remarks
S126	C001	01-05-2022 10:30:42	Apply leave	hr.apply_leave	1.0	[{intent:hr.apply_leave, confidence:1.0, domain:HR},. . .]
S126	C002	01-05-2022 10:31:58	#Annual Leave	hr.apply_leave	null	null
S126	C003	01-05-2022 10:33:20	#2022-05-06#2022-05-06#False	hr.apply_leave	null	null
S126	C004	01-05-2022 10:34:03	#Submit	hr.apply_leave	null	null

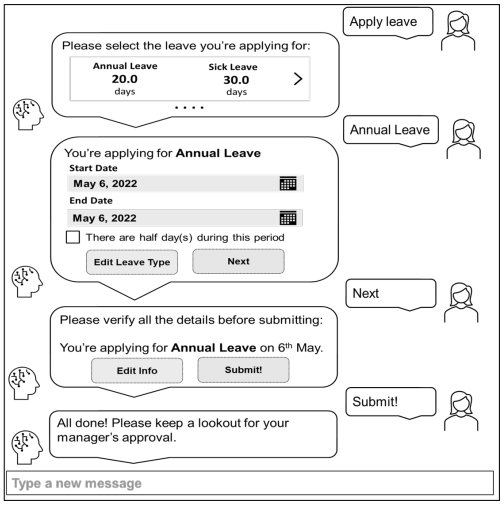

Figure 2: Example of a GC chatbot dialogue in a session

chatbot is not confident in its prediction. Based on these user-action hypotheses and intent prediction results, a framework was designed to incorporate these possibilities to extract potential False Positives.

4.1 Duplicate detection

When the user queries a similar question within a short time interval (within a chat session), it is highly possible that the chatbot response is not satisfactory. Hence, one pattern for detecting False Positives is through duplicate detection. An example of a duplicate question is shown in Figure 1, where the employee is troubleshooting for an email connectivity issue. Data preprocessing is applied to filter for conversations that are relevant user queries. For instance, user responses that are part of the GC flow and user queries that are commonly asked questions or social pleasantries are excluded. Some commonly asked questions within a chat session are 'apply leave' and 'check leave balance'. Without filtering away these queries, they would be flagged as duplicates and potential False Positives when they are not true. Other chatbot applications can come up with their own rules to filter out the irrelevant set of user queries.

The flow of the duplicate detection approach is illustrated in Figure 4. When the chatbot's response is not satisfactory, the user might paraphrase and ask again. As such, simply looking at the word overlap in the queries might be insufficient to identify duplicates. To account for paraphrasing, quora-distilbert-base [14], a pre-trained sentence transformer model, was used. This model was selected over other sentence transformer models as it was trained on the Quora Duplicate Question dataset, which contains annotations on whether two questions are duplicated or not. This is like our use case where we try to identify duplicate queries. The model maps the user query to a 768-dimensional dense vector space embedding, and the cosine similarity score of pairwise embeddings within a chat session is computed. Paired queries are flagged as duplicates and potential False Positives if the cosine similarity score is above a specified threshold. Additional post-processing rules (in Table 3) were applied to filter down the list of potential False Positives.

Labeled data was used to determine the optimal cosine similarity threshold for False Positive detection. The annotators labeled 940 query-intent records sampled across a 2-week period. The records were labeled as 1 if the intent predicted was a False Positive, and

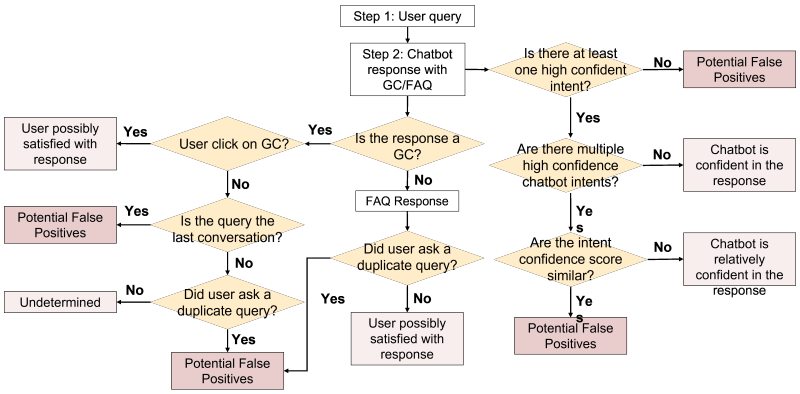

Figure 3: Flowchart for user-action hypothesis

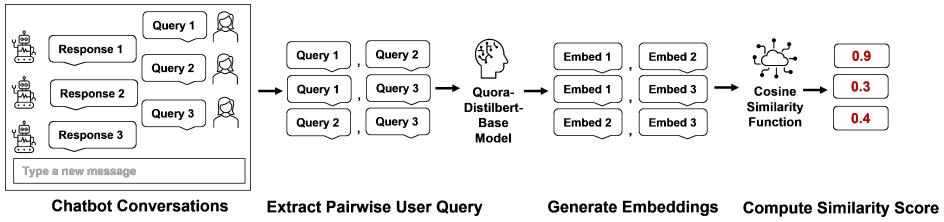

Figure 4: Flow of duplicate detection approach

Table 3: Post-processing to filter down list of False Positives

Filter Criteria	Rationale
Exclude records that are the last conversation in the chat session	Chatbot could have possibly answered user question, that's why there is no more conversation
Exclude records with maximum chatbot intent confidence > 0.9	High confidence chatbot responses are usually correct

0 otherwise. Of the 940 records, 224 were annotated as False Positives. The recall, precision, and F1-score were computed at different cosine similarity thresholds. Precision is defined as the proportion of correct False Positive intents predicted. Recall is defined as the proportion of correct False Positive intents identified. The F1-score is the harmonic mean of precision and recall. As the main objective is to prioritize the annotators' review efforts, precision was used as a key metric to determine the optimal threshold. From Figure 5, the optimal cosine similarity threshold based on precision is 0.75. Paired queries are flagged as False Positives if the similarity score is greater than 0.75. The corresponding F1-score, precision, and recall score at the 0.75 threshold are 0.181, 0.585, and 0.107, respectively.

To provide some insights into the False Positives detected, the duplicate pairs are further categorized into four groups (refer to Table 4) based on their intent. An intent is marked as 'known' if the confidence score is greater than 0.4, and 'unknown' otherwise. In addition, intents are defined as similar if they belong to the same domain. For instance, the matched intent 'hr.apply_leave' indicates that the intent belongs to the HR domain. It is important to highlight that a conversation record could be marked as a duplicate with more than one other conversation. As such, it is possible for a conversation to belong to more than one category shown in Table 4.

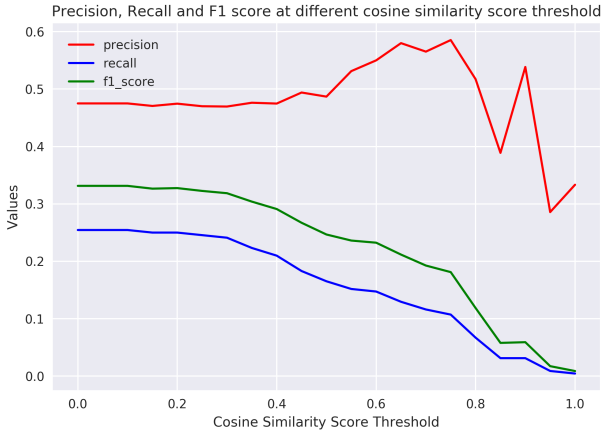

Figure 5: Precision, recall and F1-score at different cosine similarity probability threshold

Table 4: Categories of duplicates

Category	Definition	Hypothesis	Action
Category 1	Duplicates with 1 known and 1 unknown intent	The conversation with a known intent is a borderline unknown and could be a potential false positive	Annotate the record with a known intent.
Category 2	Duplicates with 2 known and same intent	Intent too generic and is not able to cater to specific query. Model learnt incorrect mapping.	Enrich intent response to specific queries (e.g., region specific responses)
Category 3	Duplicates with 2 known intents from the same domain	The intents are overlapping.	Representatives from relevant domain should review and merge intents if necessary.
Category 4	Duplicates with 2 known intents from different domain	One or both intents could be inaccurate. Highly likely to be False Positive.	Annotate both records.

Table 5: False positives detection precision by categories

Category	# FP Flagged	# True FP	FP Detection Precision
Category 1	1	0	0%
Category 2	15	8	53.3%
Category 3	7	1	14.3%
Category 4	25	19	76.0%
Total	41	24	58.5%

Categorizing the paired duplicates can help the chatbot owners further tune their intent responses and determine what relevant actions to take. For example, one can choose to merge intents that are similar and enhance the response. While other chatbots might not have the 'domain' information, one can consider utilizing their own version of grouping if available or perform topic modelling to group similar intents.

From the duplicate categories, Category 4 is expected to have the highest False Positive detection precision as two similar queries tagged to different domains have a higher potential of being False Positives. Using the same set of labeled data, records with a cosine similarity greater than 0.75 are flagged as potential False Positives and categorized into the four groups (in Table 4). The precision for each category (in Table 5) validates the hypothesis that Category 4 has the highest False Positive detection precision. Although the overall False Positive detection precision from the duplicate detection approach is already 2.46 times better than the baseline, our use case only considers the Category 4 duplicates to increase the False Positive detection precision.

4.2 Untriggered guided conversations

An untriggered Guided Conversation could also be an indication of a potential False Positive, as the user would not click on a wrongly suggested response. In such cases, the conversation could either be (1) the last conversation in the session or (2) not the last conversation in the session.

To increase the False Positive detection precision, score thresholding was performed, and records with intent confidence score lower than the threshold are flagged as potential False Positives. Since all low confidence records (confidence score less than 0.4) are already reviewed by annotators, the confidence threshold will start at 0.45. From Figure 6, it shows that the precision follows a downward trend as the confidence threshold increases. This trend is expected as a higher confidence score generally implies that the chatbot is more confident in its intent prediction and thus less likely to give a False Positive. To determine the optimal confidence score threshold, we identify the point in which there is a sharp decline in precision, which is 0.6. This is because we want to increase the False Positive detection recall while maintaining a relatively high precision.

4.3 Similar high confidence across multiple intent

The current approach of annotating all the low-confidence records is part of the least-confidence query strategy in active learning. It only looks at the confidence of the most probable intent and disregards the other intent probabilities. Using other query strategies such as margin and entropy sampling allows us to consider scenarios where the record has multiple high and similar confidence across various intents. Such scenarios could imply that the chatbot is confused and not confident in its response. In margin sampling, the differences between the top two intent probabilities are considered, whereas in entropy sampling, all the intent probabilities are used for calculation. Score thresholding was applied to determine the optimal margin and entropy threshold, and records with a margin score lower than the margin threshold or an entropy score greater than the entropy threshold are extracted as potential False Positives.

Like the intent confidence score, a higher margin score indicates that the chatbot is more confident with its prediction and thus has fewer False Positives. Figure 7 (Left) shows an overall downward

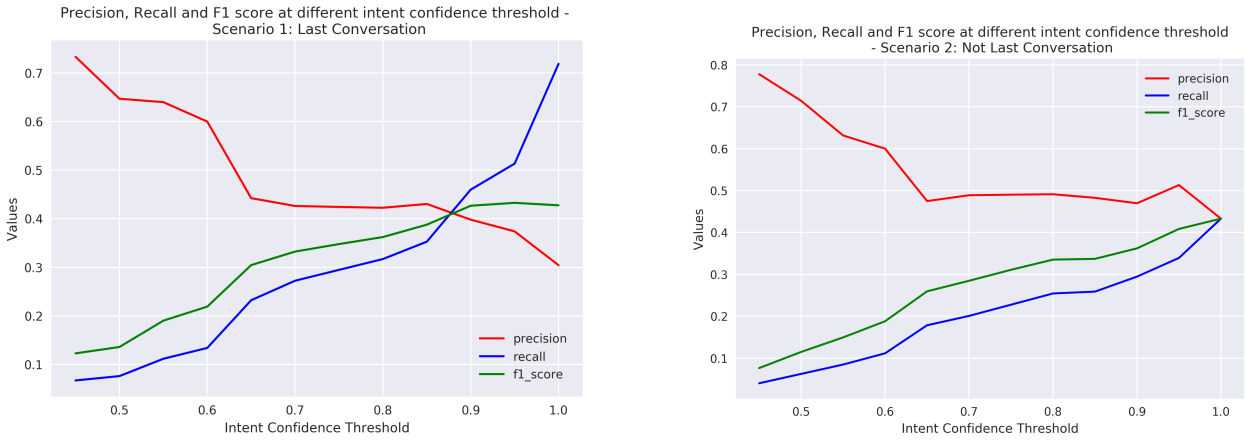

Figure 6: Precision, recall and F1-score at different intent confidence threshold for scenario 1 (Left) and scenario 2 (Right)

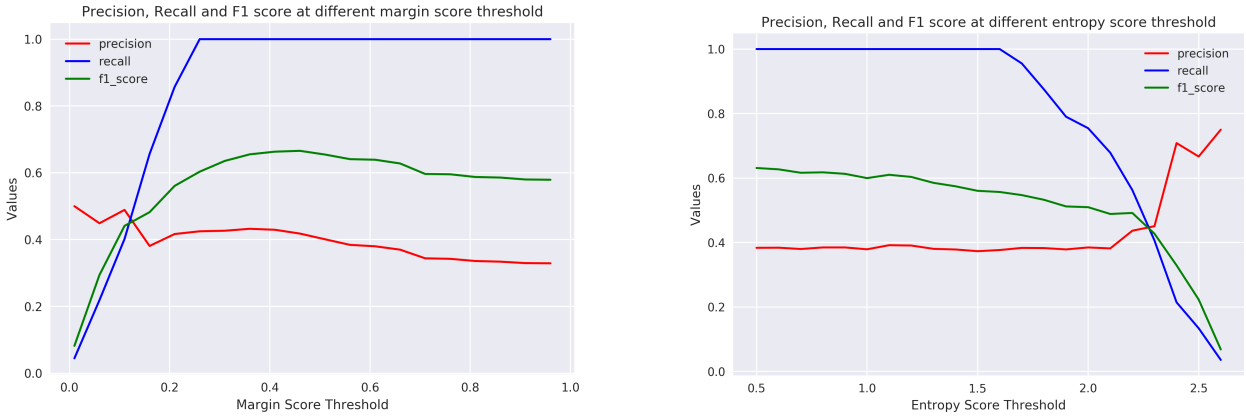

Figure 7: Precision, recall and F1-score at different margin (Left) and entropy (Right) score threshold

trend in precision, and a smaller margin gives a higher precision. To determine the optimal margin score threshold, we identify the point with a sharp decline in precision, which is 0.1. A higher entropy score, as opposed to a higher margin score, indicates that the chatbot is less confident in its prediction and thus has more False Positives. An overall upward trend in precision is observed in Figure 7 (Right) when the entropy score threshold is increased. To determine the optimal entropy score threshold, we identify the point with a sharp increase in precision, which is 2.4. Records with a margin smaller than 0.1 or entropy greater than 2.4 are flagged as potential False Positives.

4.4 Combining approaches

The approaches covered in Sections 4.1 to 4.3 are not mutually exclusive, and records could be flagged as potential False Positives under multiple approaches. To increase the False Positive detection recall while maintaining high precision, all three approaches are combined to include (1) duplicates that belong to Category 4, (2) untriggered GC responses with intent confidence less than 0.6, (3) entropy scores of intent probabilities greater than 2.4, and (4) margin scores of intent probabilities less than 0.1. The overall False Positive detection precision for the combined approaches is 67%, which is 2.81 times better than the baseline of random sampling. In addition, the False Positive detection recall is 31.7%, giving an overall F1-score of 0.430. By comparing the F1-score to the baseline F1-score of 0.384, it verifies that the False Positive detection framework performs better than random sampling. There is no baseline model used for comparison as the False Positive detection problem cannot be directly modelled. Hence, the baseline used is the annotated results from random sampling.

5 ONLINE EVALUATION

This section presents the online model performance of the False Positive detection framework. Using the combined approaches under the False Positive detection framework, a total of 1652 query-intent pair data was flagged as potential False Positives over a period of 4 months. Of which, 1174 were annotated as true False Positives. This gives an overall precision of 71.1%. The overall precision and

Table 6: Precision of False Positive Detection Approaches

Approach	# FP Flagged	# True FP	FP Detection Precision
(4.1) Duplicate Detection	68	54	79.4%
(4.2) Untriggered Guided Conversation	4	3	75.0%
(4.3) Similar high confidence across multiple intents	74	55	74.3%
(4.1 & 4.2) Duplicate Detection & Untriggered Guided Conversation	0	0	-
(4.1 & 4.3) Duplicate Detection & Similar high confidence across multiple intents	825	634	76.8%
(4.2 & 4.3) Untriggered Guided Conversation & Similar high confidence across multiple intents	681	428	62.8%
(4.1 & 4.2 & 4.3) Duplicate Detection & Untriggered Guided Conversation & Similar high confidence across multiple intents	0	0	-
Overall	1652	1174	71.1%

the precision breakdown by approach are highlighted in Table 6. The numbers shown in Table 6 are non-overlapping and can be summed up to achieve the overall numbers at the bottom of the table.

6 CONCLUSIONS

Although there have been many research efforts on improving chatbot capabilities, these are mostly focused on training strategies using annotated training data and feedback loops. Little has been done on the chatbot dialogue analysis and improvement process, which is equally as important as it allows one to improve on the chatbot's content and capability for intent recognition so that it may respond correctly to similar requests in the future. One major improvement process is the identification of False Positives. With limited to no explicit feedback data as labels, False Positive intent detection is a challenging problem to solve. To mitigate this challenge, a framework was introduced in this paper. Applying this framework to an employee chatbot's data shows that the approaches significantly outperform the random sampling baseline in terms of precision. Although some initial annotation efforts are required to determine the optimal thresholds for each approach, the subsequent annotation efforts will be more targeted and productive as the list of records flagged as False Positives gets more accurate.

The framework captures implicit feedback data using techniques such as duplicate detection to identify user behaviors that suggest dissatisfaction with the chatbot responses. Patterns such as repeated queries and untriggered guided conversations were discussed in this paper. Apart from using implicit feedback, active learning query strategies such as margin and entropy sampling were also included to capture instances where the chatbot is not confident in its intent predictions. A huge advantage of this framework is that it is scalable and labeled data is not required to train any model. One can simply validate on a smaller set of labels to finetune the approaches and determine the optimal threshold specific to their use case. Furthermore, the implicit patterns such as paraphrasing of queries and exiting the conversation after the chatbot replies should be common across chatbots. Hence, the framework can be reused across different chatbot applications. The framework was validated on multiple periods of labeled data sets for our use case, and the precision is consistent. Specifically, the combined approaches show a high False Positive detection precision of 67% and 71.1% in an offline and online test set, respectively. This approach of detecting False Positives algorithmically and accurately translates into more efficient and productive annotation efforts and, in turn, reduced model improvement lead time.

ACKNOWLEDGMENTS

We would like to thank the collaborating departments and sponsors within DBS bank: Sameer Gupta, Cynthia Ang, Lauren Li, Caroline Maheshwari, Laine Ong, Jun Sun Tjhin, Jason Tan, and other team members.

REFERENCES

[1] Enterprisersproject.com. 2022. What is digital transformation?. Retrieved January 1, 2022 from https://enterprisersproject.com/what-is-digital-transformation

[2] Asbjørn Følstad and Marita Bjaaland Skjuve. 2019. Chatbots for Customer Service: User Experience and Motivation. In Proceedings of the 1st International Conference on Conversational User Interfaces, (Dublin, Ireland). https://doi.org/10.1145/3342775.3342784

[3] Uroš Arsenijevic and Marija Jovic. 2019. Artificial Intelligence Marketing: Chatbots. In Proceedings of the International Conference on Artificial Intelligence: Applications and Innovations (IC-AIAI), (Belgrade, Serbia). https://doi.org/10.1109/IC-AIAI48757.2019.00010

[4] Mary Bates. 2019. Health Care Chatbots Are Here to Help. IEEE Pulse, 10(3), 12-14. https://doi.org/10.1109/MPULS.2019.2911816

[5] Raphael Meyer von Wolff, Sebastian Hobert, Kristin Masuch and Matthias Schumann. 2020. Chatbots at Digital Workplaces – A Grounded-Theory Approach for Surveying Application Areas and Objectives. Pacific Asia Journal of the Association for Information Systems, 12(2), Article 3. Available at https://aisel.aisnet.org/pajais/vol12/iss2/3

[6] Chayan Chakrabarti and George F. Luger. 2015. Artificial Conversations for Customer Service Chatter Bots: Architecture, Algorithms, and Evaluation Metrics. Expert Systems with Applications, 42(20), 6878-6897. https://doi.org/10.1016/j.eswa.2015.04.067

[7] Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo and Rama Akkiraju. 2018. Touch Your Heart: A Tone-aware Chatbot for Customer Care on Social Media. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, (Montreal, QC, Canada). https://doi.org/10.48550/arXiv.1803.02952

[8] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha and Rama Akkiraju. 2017. A New Chatbot for Customer Service on Social Media. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, (Denver, CO, USA). https://doi.org/10.1145/3025453.3025496

[9] Knut Kvale, Olav Alexander Sell, Stig Hodnebrog and Asbjørn Følstad. 2020. Improving Conversations: Lessons Learnt from Manual Analysis of Chatbot Dialogues. In Proceedings of the Chatbot Research and Design – Third International Workshop, CONVERSATIONS 2019, (Amsterdam, The Netherlands). https://doi.org/10.1007/978-3-030-39540-7_13

[10] Asbjørn Følstad and Cameron Taylor. 2020. Conversational Repair in Chatbots for Customer Service: The Effect of Expressing Uncertainty and Suggesting

Alternatives. In Proceedings of the Chatbot Research and Design – Third International Workshop, CONVERSATIONS 2019, (Amsterdam, The Netherlands). https://doi.org/10.1007/978-3-030-39540-7_14

[11] Zahra Ashktorab, Mohit Jain, Q. Vera Liao and Justin Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, (Glasgow, Scotland, UK). https://dl.acm.org/doi/10.1145/3290605.3300484

[12] Asbjørn Følstad, Cecilie Bertinussen Nordheim and Cato Alexander Bjørkli. 2018. What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study. In Proceedings of the 5th International Conference on Internet Science, INSCI 2018, (St. Petersburg, Russia), Cham, Switzerland: Springer (2018). https://doi.org/10.1007/978-3-030-01437-7_16

[13] Samarth Agarwal, Yang-yu Tseng, Ying Yang Lee, Xuejie Zhang and John Lu. 2021. Smart Annotation using Semantic Search for Employee Chatbot Platform in Financial Services. In Proceedings of the KDD Workshop on Machine Learning in Finance, (Virtual workshop).

[14] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, (Hong Kong, China). https://doi.org/10.48550/arXiv.1908.10084