Sentiment-Aware Dialogue Flow Discovery for Interpreting Communication Trends

Patrícia Ferreira[1,2] **Isabel Carvalho**[1,2] **Ana Alves**[1,3]

Catarina Silva[1,2] **Hugo Gonçalo Oliveira**[1,2]

[1] **CISUC, LASI,** [2] **DEI, University of Coimbra, Portugal**

[3] **Polytechnic Institute of Coimbra, Portugal**
{patriciaf,isabelc,ana,catarina,hroliv}@dei.uc.pt

Abstract

Customer-support services increasingly rely on automation, whether full or with human intervention. Despite optimising resources, this may result in mechanical protocols and lack of human interaction, thus reducing customer loyalty. Our goal is to enhance interpretability and provide guidance in communication through novel tools for easier analysis of message trends and sentiment variations. Monitoring these contributes to more informed decision-making, enabling proactive mitigation of potential issues, such as protocol deviations or customer dissatisfaction. We propose a generic approach for dialogue flow discovery that leverages clustering techniques to identify dialogue states, represented by related utterances. State transitions are further analyzed to detect prevailing sentiments. Hence, we discover sentiment-aware dialogue flows that offer an interpretability layer to artificial agents, even those based on black-boxes, ultimately increasing trustworthiness. Experimental results demonstrate the effectiveness of our approach across different dialogue datasets, covering both human-human and human-machine exchanges, applicable in task-oriented contexts but also to social media, highlighting its potential impact across various customer-support settings.

1 Introduction

Dialogue systems are increasingly pervasive, playing a crucial role in communication with customers in many companies. Monitoring and visualizing conversations produced by such systems offers a deeper comprehension of dialogue interactions, unveiling communication patterns, and providing valuable insights into the user experience. It is thus essential to ensure high-quality service. Here, the analysis of frequent dialogue flows plays an important role, as they will depict the organic evolution of interactions, enhancing human interpretability.

Obtaining dialogue flows from black-box systems, such as chatbots based on Large Language Models (LLMs) or other encoder-decoder frameworks, can be challenging due to their generative and open-domain nature. Nonetheless, the ability to represent the conversation progression and consider emotional aspects such as the sentiment of the speakers is valuable, especially in activities requiring real-time assistance from responsible agents.

We propose an approach for automatic dialogue flow discovery from a history of written dialogues, and their representation in a transition graph. We begin by grouping similar utterances into clusters, which may be seen as dialogue states. Then we represent possible paths with their respective probabilities from the beginning to the end of the dialogue.

Furthermore, we enrich the states with the average sentiment of the included utterances. This has applications in a wide range of services and products involving dialogue or customer support, including call centers, emergency services, and virtual assistantsIt also serves as an assessment tool, offering stakeholders a way to compare dialogue systems based on how they handle client requests while maintaining or improving their sentiment. Moreover, this approach can potentially identify topics that often result in negative sentiment. The main contributions of this work are summarized as:

- The proposal of a solution for the automatic discovery of dialogue flows that are adaptable to any language and domain, offering an interpretability layer to dialogue systems;
- The integration of sentiment analysis into existing/automatically generated flows, enriching interpretability with sentiment variations;
- The proposal of flow metrics for assessing (i) agents' performance based on sentiment variation, (ii) effectiveness in capturing common states, and (iii) sentiment and cluster cohesion within flows;
- A visual analysis of flows discovered from

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 274–288
September 18–20, 2024.

diverse dialogue datasets, spanning various services and types, complemented by the proposed metrics, while showcasing the proposed approach and confirming its benefits;

- A proposal for an advanced analysis layer that includes sentiment variation representation within each cluster, offering valuable insights for assessing agent performance and identifying sentiment-associated states.

The remainder of the paper is structured as follows: Section 2 reviews work related to dialogue flow discovery and sentiment analysis; Section 3 describes the proposed approach for sentiment-aware dialogue flow discovery; Section 4 clarifies the meaning of each element in the sentiment-aware dialogue flows, and describes the experimental setup, the used datasets, and the flow metrics proposed; Section 5 presents and analyses the resulting flows; Finally, Section 6 concludes the paper and provides cues for future work.

2 Related Work

The categorization of utterances in dialogue systems may help in understanding user intentions and facilitating effective interactions (Deng et al., 2023; Gonçalo Oliveira et al., 2022). Generally, utterances are classified according to user intentions (Vedula et al., 2020; Mou et al., 2022) or dialogue acts (Ribeiro et al., 2019; Liu et al., 2017), both providing valuable insights for task-oriented systems. However, the automatic classification of utterances is typically supervised and thus relies on annotated data, which is not always available. On the other hand, encoder-decoder systems, including those based on LLMs (e.g., ChatGPT[1]), do not rely on such classifications, but their flexibility comes at the cost of higher data demand and less control.

Traditional task-oriented dialogue systems are sustained by the design of flows to guide conversations towards specific goals. This entails defining specific user intentions and training phrases, and can be facilitated by tools like Google's DialogFlow[2], Microsoft Luis[3], or Rasa[4]. Automating this process involves grouping semantically-similar utterances and representing them in a vector space, towards efficient intent discovery (Hashemi et al., 2016; Park et al., 2022; Liu et al., 2021).

[1] https://chat.openai.com/
[2] https://cloud.google.com/dialogflow
[3] https://www.luis.ai/
[4] https://rasa.com/

Representing dialogue flows as transition graphs offers insights on topics and other trends (Bouraoui and Lemaire, 2017). An earlier approach (Ritter et al., 2010) for flow discovery uses Hidden Markov Models on Twitter conversations. It introduces features like clustering similar utterances, vertices for marking the beginning and end of dialogues, as well as a threshold for ignoring low-probability transitions. Towards interpretability, clusters were labelled manually. Ferreira et al. (2024) developed a similar approach with automatic labelling.

By analysing communication trends, flow discovery may assist in the design of dialogue systems. This is the main goal of Graph2Bots (Bouraoui et al., 2019), which adopts co-clustering for discovering dialogue states and transitions in human-human conversations. An alternative approach (Sastre Martinez and Nugent, 2022) clusters utterances with DBSCAN and relies on finite-state automata for discovering ranked flows, based on the frequency of question-response sequences.

Sentiment Analysis (SA) (Liu, 2015) aims to extract sentiments from texts. In dialogues, it may help in identifying situations of sentiment degradation, which may then be acted upon, e.g., through a fallback system that replaces an artificial agent by a human; or by collecting information for later retraining the human or artificial agent.

SA has been combined with other tasks, such as dialogue act recognition, which reinforce one another. For instance, detecting agreement often corresponds with the expression of the same sentiment, while transitions from negative to neutral tend to coincide with changing to a statement. Works that tackled these tasks jointly (Xu et al., 2023; Qin et al., 2020; Li et al., 2020) benefited from it, and achieved high or state-of-the-art (SOTA) performances in datasets like Mastodon (Cerisara et al., 2018). Moreover, Song et al. (2023) outperformed several SOTA methods for user satisfaction estimation in task-oriented dialogue systems by exploiting SA in a multi-task adversarial strategy.

The seemingly symbiotic relationship between SA and other tasks motivated its application to dialogue flow discovery. Yet, to the best of our knowledge, no other work has combined these tasks.

3 Proposed Approach

We propose a generic approach for automatically discovering the most common flows in a history of dialogues, while simultaneously associating sen-

timent with their transitions. It comprises three distinct steps, outlined in Figure 1:

1. **Utterance Clustering** clusters semantically similar utterances, represented by their embedding. Discovered clusters may be seen as approximations to dialogue states.

2. **Flow Discovery** computes the most frequent paths. The result is a transition graph $G(C, T)$, where nodes $c \in C$ represent dialogue states and edges $t(c_i, c_j, p_{ij}) \in T$ represent transitions. The latter are weighted according to their probability, computed as in Equation 1, where $|t(c_i, c_j)|$ represents the number of utterances in c_j that immediately follow a utterance in c_i.

$$p_{ab} = \frac{|t(c_a, c_b)|}{\sum_{x \in C} |t(c_a, c_x|} \quad (1)$$

3. **Sentiment Classification** enriches G with sentiment information. When sentiment is not available with the data, each utterance's sentiment can be determined by an external tool for the purpose. This enables the computation of the most predominant sentiment in each state and transition.

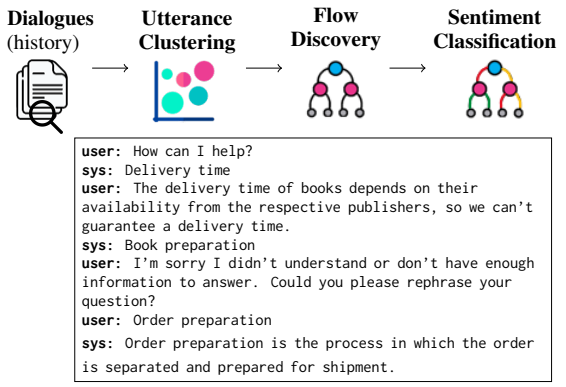

Figure 1: Overview of the proposed approach and an illustrative dialogue between a customer (user) and an artificial agent (sys).

The proposed approach can be applied to any collection of written dialogues, ideally with two speakers, but of any type, in any language, or domain, as long as utterances are provided in sequence and speakers are identified. The graph G can be visually represented, thereby enhancing human interpretation. The sentiment may be visually represented with different colours for each edge, such as green, yellow, and red, representing predominantly positive, neutral, or negative sentiments, respectively (as shown on the right-hand side of Figure 1). Sentiment-aware dialogue flows can be useful in various scenarios, including:

Identifying communication trends i.e., the discovery of flows from any type of dialogue promotes the identification of common and/or undesired topics or transitions, which can be used to improve the agent, e.g., by changing intents, reviewing protocols, or adjusting human resources;

Interpreting black-box dialogue systems i.e., the discovery of flows in human-machine dialogues adds an interpretability layer that increases understanding of the agent and promotes the identification of issues. Potential strategies for addressing such issues may include retraining the agent or implementing additional rules;

Planning and developing dialogue systems i.e., the analysis of human-human dialogues towards the identification of potential dialogue states and representative words or sentences, valuable to the agent's development process.

In any scenario, the dialogue collection should be as comprehensive as possible and, ideally, cover all relevant intents. The set of applications attests to the versatility of the approach. Still, in this paper, we focus on the interaction between a customer and an agent, where the ability to understand and efficiently manage interactions is essential for improving the quality of service and, consequently, customer satisfaction.

4 Experimentation

In order to confirm the applicability of the sentiment-aware dialogue flows, extensive experimentation was conducted. This involved the implementation of each step of the proposed approach, introduced in Figure 1, with adequate tools, as well as the application to a range of dialogue datasets. This section details the implementation of the underlying processes but, before delving into the previous steps, we provide some clarifications on the visual notation used throughout the paper, aided by the illustrative diagram in Figure 2.

The diagram (G) showcases the ideal scenario, in which an agent successfully manages to switch the customer's sentiment from negative, at the Start Of the Dialogue (SOD), to positive, by the End Of the Dialogue (EOD). SOD and EOD are represented by specific nodes, which can be seen as states, represented as yellow boxes. The others

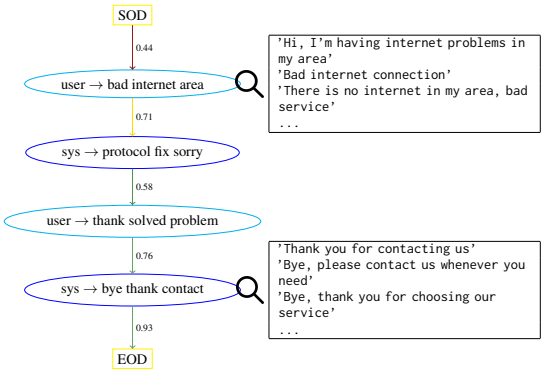

Figure 2: Example of a sentiment-aware dialogue flow showcasing an ideal scenario.

correspond to the discovered clusters (C) and are represented by ellipses labelled with representative words in their utterances. States representing clusters by different speakers are also differentiated. In G, user clusters are coloured in light blue and agent clusters in dark blue. The diagram is complemented with examples of clustered utterances, on the right-hand side.

Edges represent transitions (T) between clusters and have an associated weight, corresponding to their probability. For instance, G shows a 58% probability of moving from `sys→protocol fix sorry` to `user→thank solved problem`. The sum of all probabilities of T originating from the same cluster is 1. Nevertheless, in order to simplify the flow, a threshold can be applied for ignoring low-probability transitions, as carried out in this example. The colour of each transition represents the average sentiment within the destination cluster. Red corresponds to a more negative sentiment, green to a positive, and yellow to a neutral sentiment. For EOD, however, transitions represent the sentiment of the origin cluster, thus making the sentiment in the final interactions clearer and contributing to more immediate conclusions.

4.1 Datasets

The proposed approach was applied to five different dialogue datasets, covering different channels (social media, chat, telephone), types of dialogue (task-oriented, open) and agent (human, machine), domains (tourism, telecommunications, retail, open) and languages (English and Portuguese). Specifically, the following datasets were used:

EmoWOZ (Feng et al., 2022) a public dataset of task-oriented dialogues that extends MultiWOZ (Budzianowski et al., 2018), thus covering multi-domains related to tourism. EmoWOZ's additionally has emotions assigned to utterances, including valence, translatable to a polarity (positive, neutral, negative).

TwitterDialogueSAPT (TDSAPT) (Carvalho et al., 2023) a public dataset of customer-support dialogues in Portuguese, extracted from Twitter, with entities (i.e., accounts) in the domains of Telecommunications, Television, Healthcare, eCommerce, and Finance, where utterances have manually-annotated sentiment. We adopted the original approach for extending this dataset with more dialogues from the same entities in the same timeline (April–May, November–December 2022).

TelecomSAPT transcriptions of customer-support dialogues, sampled from two months in the call center of a Portuguese Telecommunications company, with manually-labelled sentiment.

RetailPT a collection of customer-support dialogues of a Portuguese retail company, collected during a seasonal campaign that lasted 2.5 months (July–September 2023). Dialogues are between human customers and a proprietary Retrieval Augmented Generation system based on fine-tuning an optimised version of Quokka[5][6].

Mastodon (Cerisara et al., 2018) a public dataset of dialogues extracted from the Mastodon social network, particularly from the octodon.social instance, with manually-annotated sentiment. These are open-domain conversations between two users and, as such, do not involve a service.

For some datasets, we could get the polarity of the utterances from available annotations. This was, however, not the case of RetailPT and the extension of TDSAPT, which employed a classifier fine-tuned in similar data (see Section 4.2).

Since the labels in TDSAPT were binary (negative and non-negative), we binarised the labels of all datasets, which still enabled the identification of negative transitions, the most problematic.

Table 1 describes the datasets according to channel (Chat, Phone, Social Media - SM) type of dialogue (Task Oriented - TO; Open) type of agents, domain, language (English - EN; Portuguese - PT), and number of dialogues.

[5] hf.co/automaise/quokka-7b

[6] Both TelecomSAPT and RetailPT were gently transferred to our team in the scope of projects with the industry, but are proprietary and cannot be publicly released.

Dataset	Channel	Type	Agent	Domain	Lang	#Dialogs
EmoWOZ	Chat	TO	Human	Tourism	EN	10,253
RetailPT	Chat	TO	Machine	Retail	PT	3,317
TelecomSAPT	Phone	TO	Machine	Telecom	PT	1,000
TDSAPT	SM	TO	Human	Several	PT	2,575
Mastodon	SM	Open	Human	Open	EN	535

Table 1: Brief description of each dataset, including channel, type of dialogue, type of agents, domain, language, and number of dialogues.

Table 2 presents the number of utterances in each dataset, the sentiment distribution (negative and non-negative) and informs on how the sentiment labels were obtained: in the data (D), automatic (A) by a supervised model, converted (C).

Dataset	# Utterances	% Neg	% Non-Neg	Source
EmoWOZ	140,801	1.57	98.43	C
RetailPT	19,098	28.79	71.21	A
TelecomSAPT	5,312	18.39	81.61	D
TDSAPT	5,966	36.15	63.85	D+A
Mastodon	2,205	31.61	68.39	D

Table 2: Analysis of the sentiment distribution in each dataset, including the source of sentiment labelling.

Tables 1 and 2 confirm the diversity of covered scenarios. They encompass various channels, dialogue types, agents, domains, and languages, attesting to the generalisation potential of the proposed approach. Datasets differ in size and prevalence of negative sentiment, spanning from as low as 1.6% of utterances in EmoWOZ to 36% in TDSAPT.

4.2 Implementation

Utterance embeddings were obtained from sentence transformers available in the HuggingFace Hub. Different models were used for English[7] and Portuguese[8], both representing textual sequences in 384-dimension vectors.

Clustering was performed with the K-means method, as available in scikit-learn[9]. For each dataset, the number of clusters was optimised for maximizing the Silhouette score (Rousseeuw, 1987), which evaluates the cohesion and separation of formed groups. This relied on Optuna[10], considering a range of 3–10 clusters for each speaker.

For the textual labels of the clusters, a document is created for each cluster, with its concatenated utterances. Using the same models as in the clustering step, the label resulted from the most frequent keyword for each cluster, obtained with KeyBERT (Grootendorst, 2020), considering a range of [1–3]-grams, and after removing stopwords in the NLTK (Bird and Loper, 2004) lists.

The sentiment of unlabeled utterances in TDSAPT and RetailPT was classified with a BERT model pretrained for Portuguese (Souza et al., 2020), fine-tuned for identifying negative and non-negative sentiments in Portuguese dialogues, in a similar fashion to the best model in related work (Carvalho et al., 2023). The main difference was the fine-tuning datasets, selected for sharing more similarities with the data to classify: in the extension of TDSAPT, the model was fine-tuned in the original dialogues of TDSAPT, with a 75% F1-score on it, whereas in RetailPT it was fine-tuned in TelecomSAPT, with a 74% F1-score on the former.

Finally, for representing the sentiment in each cluster, we compute the average sentiment in all its utterances. If the average sentiment is low ($<$0.4), high ($>$0.6), or in-between, we colour the incoming transitions in red, green or yellow, respectively. As the range of values associated with green and red is larger, we further define colour gradients: if the average sentiment is closer to 0.0 or 1.0, the corresponding colour gets darker. We recall that, as an average, this value may not represent the sentiment of all the utterances in each cluster. Hence, we propose a second, more in-depth analysis that includes the standard deviation (STD) of the sentiment in each cluster. Specifically, we compute: (i) the average sentiment (AVG); (ii) the sentiment at the highest deviation point (AVG+STD and assigning the corresponding colour); and (iii) the sentiment at the lowest deviation point (AVG-STD). This is considered in the graphical visualisation by adding a three-layered box to each cluster, with a larger middle layer coloured with the average sentiment, and the others with the sentiment at the lowest (left) and highest (right) deviation points. Some resulting dialogue flows are presented in Section 5.

4.3 Flow metrics

The discovered flows contribute to faster analysis of trends in the underlying dialogue datasets, but comparing flows from different datasets can still be subjective. To complement the analysis and make the comparison more straightforward, we designed objective metrics, computed directly from the flows. They capture the following aspects: (i) the agents' performance based on the sentiment throughout the dialogue flow; (ii) how well the clusters represent the dataset based on the proportion of dismissed utterances; (iii) the flow's cohesion regarding sen-

[7] hf.co/sentence-transformers/all-MiniLM-L6-v2
[8] https://tinyurl.com/2fcwpuz7
[9] https://tinyurl.com/4ymet8ff
[10] optuna.org/

timent and its clusters. The computed metrics are described ahead, with customer support in mind.

$EOD\_$**:** proportion of utterances that reach EOD with a negative sentiment. This can be applied to other sentiment levels but negativity is the one that should be mitigated;

$\Delta Sentiment$**:** difference between the sentiment of the utterances at the end of the dialogue and of those at the start.

$SOD \nrightarrow EOD$**:** proportion of utterances from all speakers that were dismissed in the flow, i.e., those that took paths with a probability lower than the set threshold, ending up not represented;

Flow Cluster Cohesion (FCC): average Silhouette score of the clusters;

Flow Sentiment Cohesion (FSC): average standard deviation of the sentiment at each cluster;

Average Initial Sentiment (AIS): average sentiment of each cluster with an incoming transition from SOD. As opposed to the values considered in $\Delta Sentiment$, this is calculated by cluster (i.e., each contains the average sentiment of the utterances within) and not by utterance.

Average Final Sentiment (AFS): average sentiment of each cluster with outgoing transitions to EOD. As opposed to $\Delta Sentiment$, this is computed by cluster, not by utterance.

An analysis of these metrics should be enough to get insights on the performance of the agent(s). Ideally, it would present (i) a low $EOD\_$, i.e., managed to avoid negative sentiment, (ii) a positive $\Delta Sentiment$, i.e., sentiment improved throughout the flow, (iii) a low $SOD \nrightarrow EOD$, i.e., most utterances were represented, (iv) a high FCC, i.e., data fits the clusters well, and (v) a low FSC, i.e., sentiment at each cluster does not deviate much. Finally, AIS and AFS should be analysed together as the latter should be higher than the former, i.e., sentiment at the cluster level should improve throughout the flow. The next section reports on applying these metrics to the considered datasets.

5 Results and Discussion

Flows were discovered from every considered dataset and the designed metrics were computed as well. Together, they provide insights into interpretability, communication trends, and limitations of the agents, among others. This section discusses some of the discovered flows and reports on the metrics computed for all. Due to lack of space, we do not present the flows for all datasets, but include them in the Appendix A.

The dialogue flow for RetailPT data is presented in Figure 3[11]. Various interactions between the user and the (artificial) agent can be observed. We immediately note that the first interaction of the agent (SOD's outgoing transition), is always the same, with probability 1.0. The label of the initial cluster suggests an offer of assistance, which is confirmed by the data: in fact, all dialogues start with the `How can I help?` utterance.

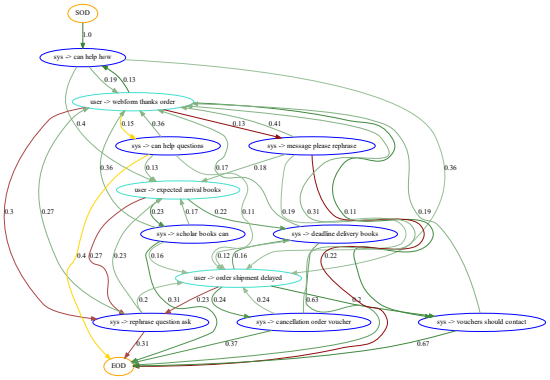

Figure 3: Sentiment-aware dialogue flow discovered for RetailPT.

With the help of the labels, we see that this interaction is followed by a user message: thanking for the order; querying about the arrival of the books; or informing on a shipment delay. As the probabilities of each transition from the `can help how` state do not sum up to 1.0, there is at least one low-probability transition (i.e., $p < 0.1$) not represented. Afterwards, the agent replies and, in some cases, asks the user to rephrase the question. Interactions continue until the EOD, which marks the end of the conversation.

In this case, non-negative sentiments (i.e., shades of green) predominate. Additionally, there are edges with neutral colours leading to the `can help questions` state and red edges associated with negative sentiments in the `rephrase question ask` and `message please rephrase` states. We may deduce that the agent is failing to process the requests, which could potentially increase the user's frustration as some end the conversation afterwards.

[11] RetailPT data is in Portuguese. For an easier interpretation by the readers, cluster labels were translated to English.

This negative sentiment could potentially be mitigated by retraining the agent to better handle the queries that lead to those clusters.

It is important to note that the colour of the edges represents the average sentiment of utterances in each cluster, which may not fully capture the sentiment of the entire cluster. Hence, we created an advanced analysis layer, shown in Figure 4, which considers the sentiment's standard deviation for each cluster via a three-layered cluster.

In the `can help questions` state, the average sentiment is represented by the colour yellow (i.e., neutral). Its left layer (red) represents the sentiment at its lower deviation value and its right layer (green) represents the sentiment at its highest deviation value. In this case, the average does not accurately represent the sentiment within that cluster as it also includes strong negative and positive values (i.e., deep shades of red and green).

In states such as `scholar book can` or `vouchers should contact`, there is minimal sentiment deviation, as each layer of the node appears uniformly green, suggesting that the sentiment within the utterances of underlying clusters is accurately represented by their average.

Table 3 reports on metrics computed for the utterances' transitions and their sentiment. We recall that these can be used to evaluate an agent's performance and how well the flow captures common states, i.e., represents most utterances.

EmoWOZ has $EOD_{-} = 0$, meaning no dialogue ends with negative sentiment. Moreover, it has the highest $SOD \nrightarrow EOD$, meaning that, with the applied threshold (0.1), most utterances are lost along the way. As this is the largest dataset (seven times larger than RetailPT) it makes sense that it would be challenging to represent each utterance in it. Sentiment variation is the lowest for this dataset.

TelecomSAPT has the highest EOD_{-}, meaning it is the dataset that mostly finished with negative sentiment, followed by RetailPT. This means that the involved (artificial) agents could benefit from an in-depth analysis, possibly culminating in reviewing and/or retraining. These are also the only datasets with a negative sentiment variation, i.e., by the end of the dialogue, sentiment gets lower. They also show high $SOD \nrightarrow EOD$, as does Mastodon, meaning these three datasets lose over half of their utterances throughout the flow.

Mastodon and TDSAPT show intermediate values overall and the latter has the lowest $SOD \nrightarrow EOD$, meaning that more than half the utterances are represented in the flow. Both datasets have a positive sentiment variation, suggesting an improvement by the end of the conversation.

In both cases, it is not easy to speculate more. Mastodon has social media dialogues, where sentiment can flow, without clear negative consequences as in customer-support. Moreover, TDSAPT includes dialogues with a broad range of entities, and would benefit from a future analysis of the flows for each, independently.

Dataset	EOD_{-}	$\Delta Sentiment$	$SOD \nrightarrow EOD$
EmoWOZ	0.0	0.02	0.83
RetailPT	0.25	-0.28	0.63
TelecomSAPT	0.34	-0.06	0.55
Mastodon	0.08	0.18	0.65
TDSAPT	0.12	0.06	0.43

Table 3: Evaluation metrics for assessing agents' performance and flow's ability to capture common states.

Table 4 presents metrics for assessing the cohesion of flows regarding sentiment and clusters. In EmoWOZ no dialogue ends with a negative sentiment (1.00 AFS). It has also the lowest FSC,i.e., sentiment does not vary much within each cluster.

RetailPT has the highest AIS, however, AFS suggests that sentiment gets worse by the end of the dialogues. It has also the highest FCC, meaning that the data is well-fitted to the clusters.

Mastodon has the highest FSC, meaning that, contrary to EmoWOZ, sentiment diverges considerably within each cluster. However, AIS and AFS suggest that it increases by the end of the dialogue. It also presents the lowest FCC, meaning that data may not be well-fitted to the clusters, which aligns with the high divergence of sentiment within them.

TelecomSAPT and TDSAPT display intermediate results in flow cohesion and variation of sentiment within clusters. However, whereas the former's AIS and AFS suggest sentiment across the dialogues is predominantly positive, for the latter, AIS and AFS have the lowest values, suggesting a more neutral sentiment. For TDSAPT, the difference between AIS and AFS is low, as is the $\Delta Sentiment$, but in different directions. The former value should be more accurate as it computes the variation by utterance instead of cluster.

Dataset	**FCC**	**AIS**	**AFS**	**FSC**
EmoWOZ	0.11	0.96	1.00	0.12
RetailPT	0.46	1.00	0.67	0.29
TelecomSAPT	0.26	0.82	0.71	0.26
Mastodon	0.04	0.68	0.71	0.41
TDSAPT	0.14	0.57	0.55	0.31

Table 4: Flow cohesion metrics for considered datasets.

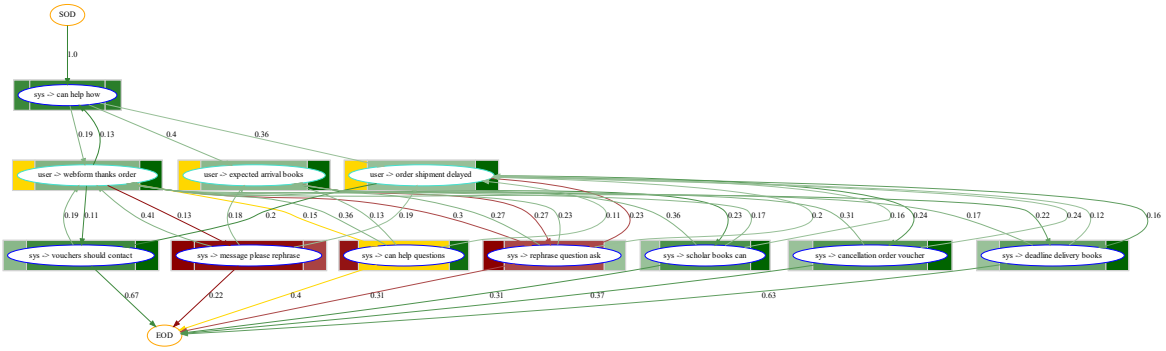

Figure 4: Sentiment-aware dialogue flow with standard deviation discovered for RetailPT data.

Finally, two factors could impact the discovery of sentiment-aware dialogue flows and, thus, their analyses: (i) the sentiment classifier, and (ii) the flow discovery process, including the clustering and labelling methods. The low performance of any of these can cause a chain reaction, decreasing the quality of the final analysis. As such, performance at each level should always be considered.

6 Conclusion

Technological advances have increased reliance on Artificial Intelligence, including for customer-support services. While efficient cost-wise, customers can tell they are interacting with an artificial agent or a human following a mechanical protocol, and this degrades their interaction and deteriorates the customers' loyalty. Our goal is to mitigate that by providing additional interpretability, also contributing to increased trustworthiness.

We proposed a novel approach for automatically discovering the most common flows in a history of dialogues, while considering the sentiment. These are useful for various applications, from identifying communication trends to interpreting black-box dialogue systems, and contribute to uncovering the triggers of problematic situations.

Our solution is independent of domain and language, and does not require dialogues labelled with intents or acts. Its implementation enabled the discovery of flows from a diverse set of dialogue datasets, out of which interesting insights were gathered, also with the help of computed metrics. For instance, in dialogues with artificial agents (RetailPT, TelecomSAPT), sentiment gets worse throughout the flow. The automation of such agents results in more mechanical answers and, thus, more cohesive clusters (FCC), when compared to other datasets. Mastodon and TDSAPT were collected from social media and cover multiple domains, which contributes to a higher variation of sentiment (FSC). Metrics also reveal that, with the parameters set (i.e., probability threshold of 0.1 and maximum 10 clusters for speaker), a large portion of utterances is lost in the flow discovery process. These regard low-probability transitions, but may degenerate interpretation, especially for large datasets as EmoWOZ. Yet, the alternative would be either to: reduce the number of clusters, with an impact on cohesion; or increase both the number of clusters and the threshold, with a negative impact in interpretability. Therefore, we plan to test alternative implementations and analyze their impact on the previous, including clustering and labelling methods, and sentiment classification, where new trends (Zhang et al., 2023) can be explored. The computation of more metrics should also be considered, e.g., for assessing the coverage of discovered flows in unseen dialogues from the same domain. Finally, towards stronger conclusions, flows should be discovered from additional datasets.

Another focus will be on flow visualization. While moving away from a graph-based model is unlikely, we consider integrating additional elements (e.g., reflecting the number of utterances in the node's size) and interactivity, towards improved interpretability (e.g., selecting the best threshold; highlighting the path taken in a specific dialogue).

Acknowledgements

This work was supported by: the project POWER (POCI-01-0247-FEDER-070365), co-financed by the European Regional Development Fund (FEDER), through Portugal 2020 (PT2020), and by the Competitiveness and Internationalization Operational Programme (COMPETE 2020); the Portuguese Recovery and Resilience Plan through project C645008882-00000055, Center for Responsible AI; and by national funds through

FCT, within the scope of the project CISUC (UID/CEC/00326/2020).

References

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Jean Léon Bouraoui, Sonia Le Meitour, Romain Carbou, Lina M Rojas Barahona, and Vincent Lemaire. 2019. Graph2bots, unsupervised assistance for designing chatbots. In *Procs. 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 114–117. ACL.

Jean-Leon Bouraoui and Vincent Lemaire. 2017. Cluster-based graphs for conceiving dialog systems. In *Procs ECML-PKDD 2017 Workshop on Interactions between Data Mining and Natural Language Processing*. CEUR-WS.org.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Isabel Carvalho, Hugo Gonçalo Oliveira, and Catarina Silva. 2023. The importance of context for sentiment analysis in dialogues. *IEEE Access*, 11:86088–86103.

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T Le. 2018. Multi-task dialog act and sentiment recognition on Mastodon. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 745–754.

Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023. A survey on proactive dialogue systems: Problems, methods, and prospects. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6583–6591. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gasic. 2022. EmoWOZ: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4096–4113, Marseille, France. European Language Resources Association.

Patrícia Ferreira, Daniel Martins, Ana Alves, Catarina Silva, and Hugo Gonçalo Oliveira. 2024. Unsupervised flow discovery from task-oriented dialogues. *arXiv preprint arXiv:2405.01403*.

Hugo Gonçalo Oliveira, Patrícia Ferreira, Daniel Martins, Catarina Silva, and Ana Alves. 2022. A Brief Survey of Textual Dialogue Corpora. In *Proceedings of the 13th Language Resources and Evaluation Conference*, LREC 2022, pages 1264–1274, Marseille, France. ELRA.

Maarten Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERT. 10.5281/zenodo.4461265.

Homa B Hashemi, Amir Asiaee, and Reiner Kraft. 2016. Query intent detection using convolutional neural networks. In *International conference on web search and data mining, workshop on query understanding*.

Jingye Li, Hao Fei, and Donghong Ji. 2020. Modeling Local Contexts for Joint Dialogue Act Recognition and Sentiment Classification with Bi-channel Dynamic Convolutions. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 616–626, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bing Liu. 2015. *Sentiment Analysis Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Pengfei Liu, Youzhang Ning, King Keung Wu, Kun Li, and Helen Meng. 2021. Open intent discovery through unsupervised semantic clustering and dependency parsing. *arXiv preprint arXiv:2104.12114*.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in dnn framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178.

Yutao Mou, Keqing He, Yanan Wu, Zhiyuan Zeng, Hong Xu, Huixing Jiang, Wei Wu, and Weiran Xu. 2022. Disentangled knowledge transfer for ood intent discovery with unified contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 46–53.

Jeiyoon Park, Yoonna Jang, Chanhee Lee, and Heuiseok Lim. 2022. Analysis of utterance embeddings and clustering methods related to intent induction for task-oriented dialogue. *arXiv preprint arXiv:2212.02021*.

Libo Qin, Zhouyang Li, Wanxiang Che, Minheng Ni, and Ting Liu. 2020. Co-GAT: A Co-Interactive Graph Attention Network for Joint Dialog Act Recognition and Sentiment Classification. *CoRR*, abs/2012.13260.

Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2019. Deep dialog act recognition using multiple token, segment, and context information representations. *Journal of Artificial Intelligence Research*, 66:861–899.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies - North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Javier Miguel Sastre Martinez and Aisling Nugent. 2022. Inferring ranked dialog flows from human-to-human conversations. In *Procs. 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 312–324, Edinburgh, UK. ACL.

Kaisong Song, Yangyang Kang, Jiawei Liu, Xurui Li, Changlong Sun, and Xiaozhong Liu. 2023. A speaker turn-aware multi-task adversarial network for joint user satisfaction estimation and sentiment analysis. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, pages 13582–13590.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS 2020)*, volume 12319 of *LNCS*, pages 403–417. Springer.

Nikhita Vedula, Nedim Lipka, Pranav Maneriker, and Srinivasan Parthasarathy. 2020. Open intent extraction from natural language interactions. In *Proceedings of The Web Conference 2020*, pages 2009–2020.

Yujun Xu, Enguang Yao, Chaoyue Liu, Qidong Liu, and Mingliang Xu. 2023. A novel ensemble model with two-stage learning for joint dialog act recognition and sentiment classification. *Pattern Recognition Letters*, 165:77–83.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

A Application to diverse data

In the following sections, we showcase the application of our approach on the five datasets presented: EmoWOZ, RetailPT, TelecomSAPT, TwitterDialogueSAPT, and Mastodon.

A.1 EmoWOZ

EmoWOZ (Feng et al., 2022) is a public dataset of task-oriented dialogues that extends MultiWOZ (Budzianowski et al., 2018), thus covering multi-domains related to tourism. It is the largest dataset covered in this work but also the one with the lowest percentage of negative utterances. It is also the only dataset where sentiment was converted as it is labelled for emotion. Figures 5 and 6 present the two sentiment-aware dialogue flows discovered for this dataset, with the latter presenting the sentiment standard deviation.

A.2 RetailPT

RetailPT is a collection of customer-support dialogues of a Portuguese retail company. Dialogues are between human customers and a proprietary Retrieval Augmented Generation system. It is the second largest dataset covered in this work and the only one in which sentiment analysis was fully automatic, by a supervised model. Figures 7 and 8 present the two sentiment-aware dialogue flows discovered for this dataset, with the latter presenting the sentiment standard deviation.

A.3 TelecomSAPT

TelecomSAPT contains transcriptions of customer-support dialogues, sampled from the call center of a Portuguese Telecommunications company, with manually-labelled sentiment. It is one of the smallest datasets covered in this work and the only one with a voice channel. Figures 9 and 10 present the two sentiment-aware dialogue flows discovered for this dataset, with the latter presenting the sentiment standard deviation.

A.4 TwitterDialogueSAPT

TwitterDialogueSAPT (TDSAPT) (Carvalho et al., 2023) is a public dataset of customer-support dialogues in Portuguese, extracted from the social network Twitter, covering accounts in multiple domains, where utterances have manually-annotated sentiment. This dataset was extended for this work, and sentiment analysis was performed automatically by a supervised model for the new utterances. Figures 11 and 12 present the two sentiment-aware dialogue flows discovered for this dataset, with the latter presenting the sentiment standard deviation.

A.5 Mastodon

Mastodon (Cerisara et al., 2018) is a public dataset of dialogues extracted from the Mastodon social network, particularly from the octodon.social instance, with manually-annotated sentiment. These are open-domain conversations between two users and, as such, do not involve a service. This is the smallest dataset covered by our work and the only one with a fully open domain and type of dialogue. Figures 13 and 14 present the two sentiment-aware dialogue flows discovered for

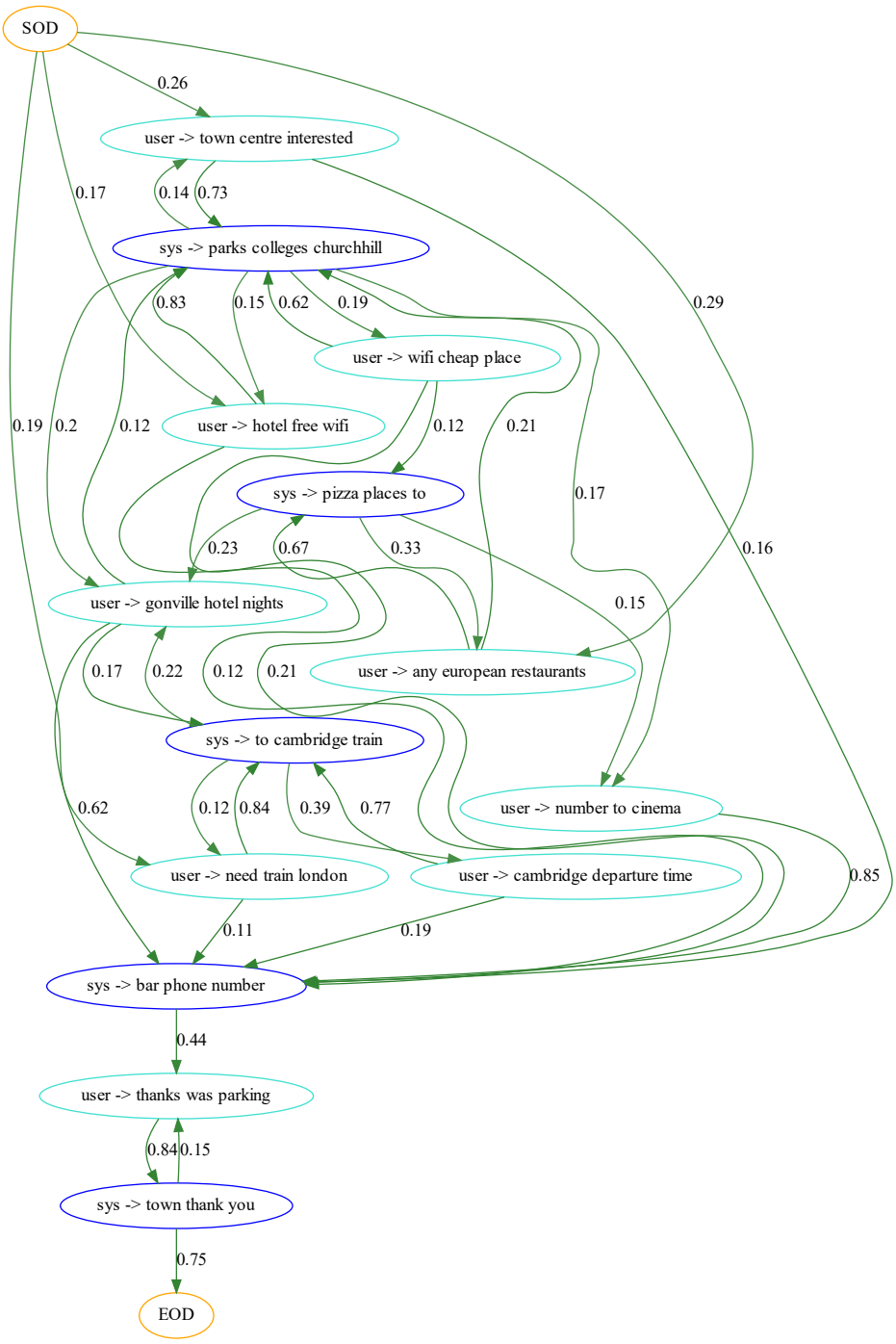

Figure 5: Sentiment-aware dialogue flow discovered for EmoWOZ data

this dataset, with the latter presenting the sentiment standard deviation.

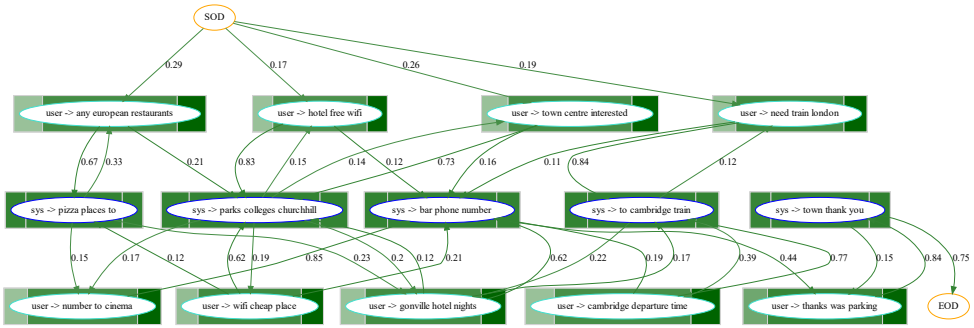

Figure 6: Sentiment-aware dialogue flow with standard deviation discovered for EmoWOZ data

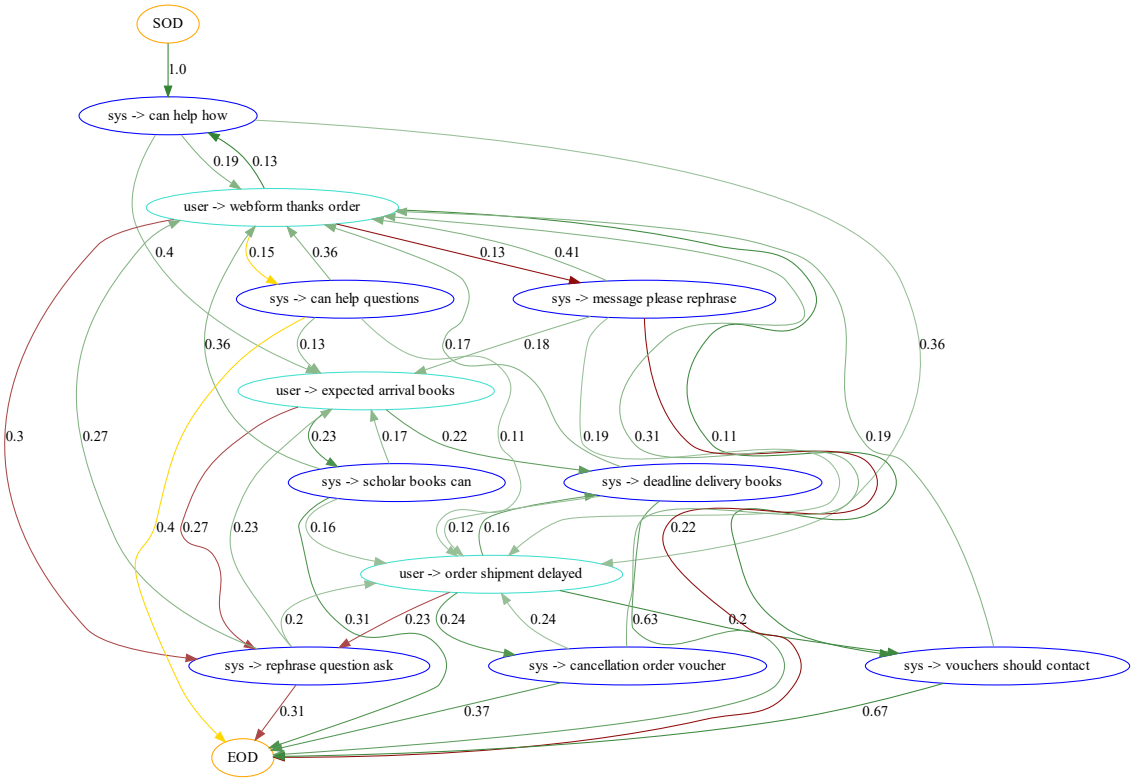

Figure 7: Sentiment-aware dialogue flow discovered for RetailPT data

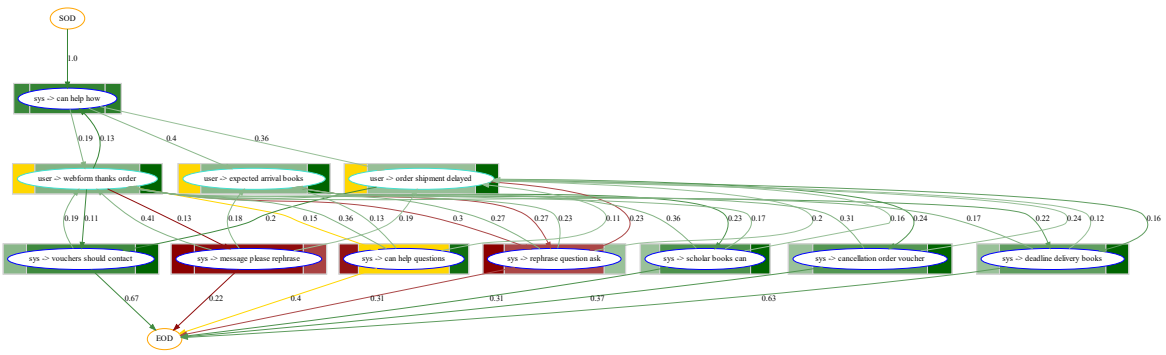

Figure 8: Sentiment-aware dialogue flow with standard deviation discovered for RetailPT data

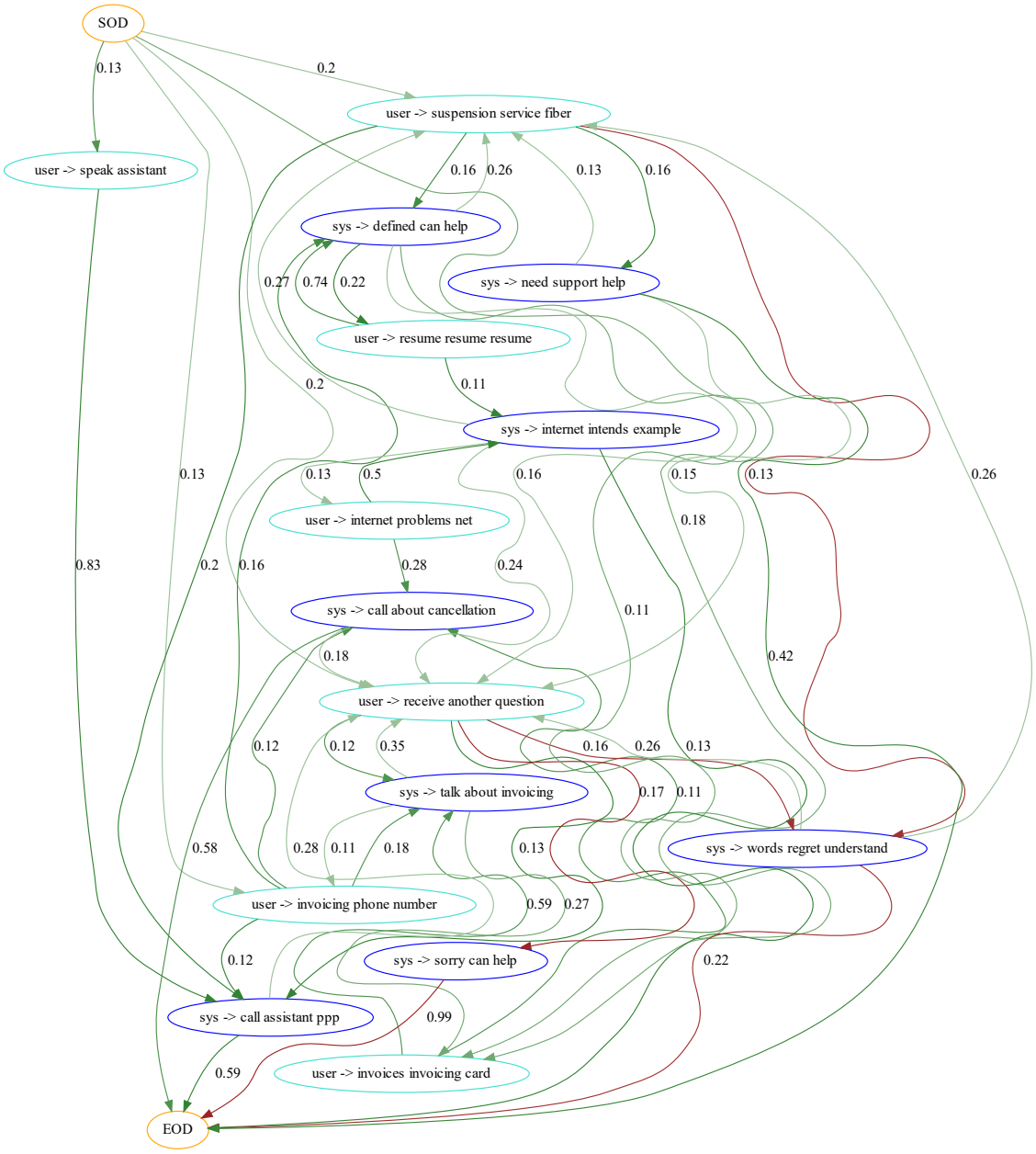

Figure 9: Sentiment-aware dialogue flow discovered for TelecomSAPT data

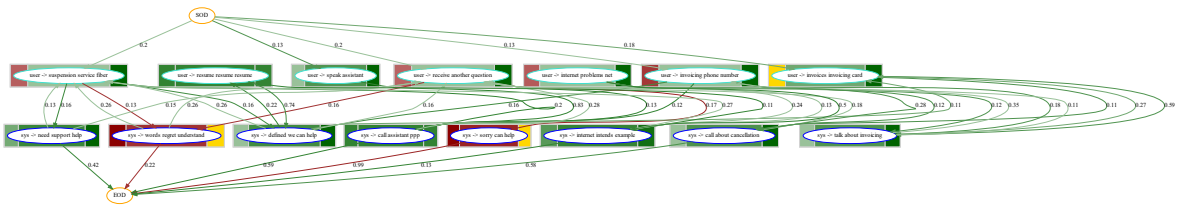

Figure 10: Sentiment-aware dialogue flow with standard deviation discovered for TelecomSAPT data

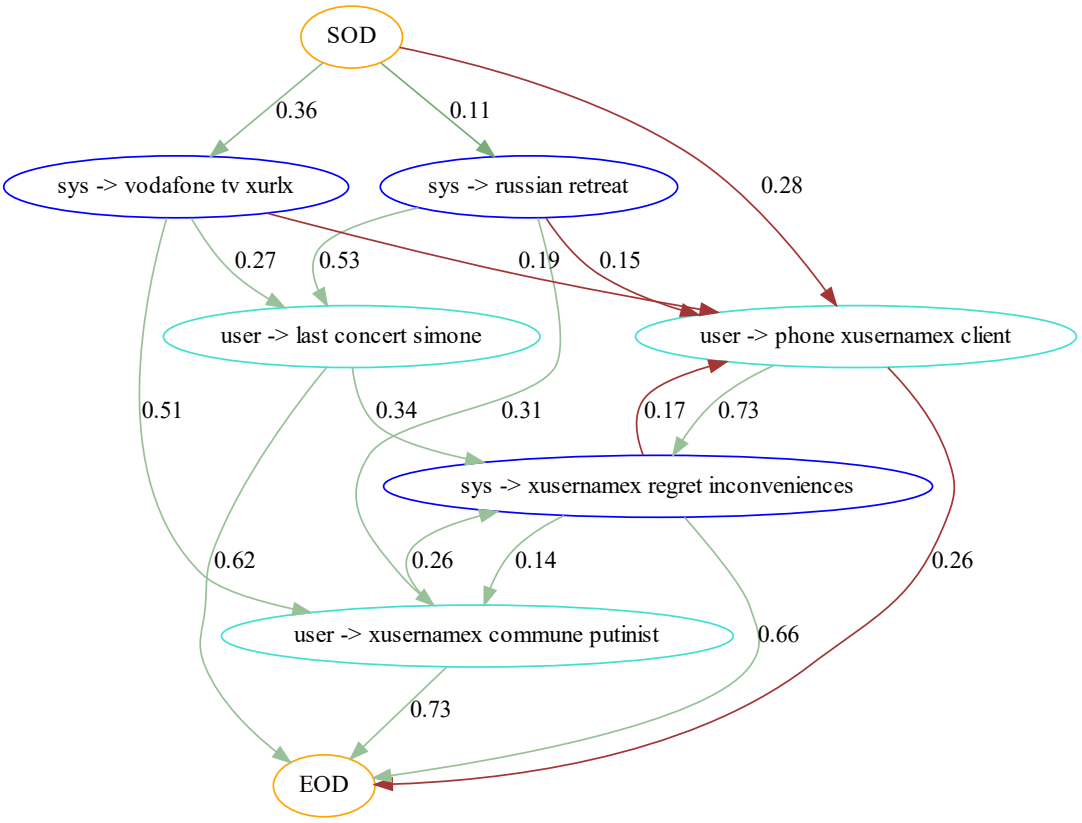

Figure 11: Sentiment-aware dialogue flow discovered for TDSAPT data

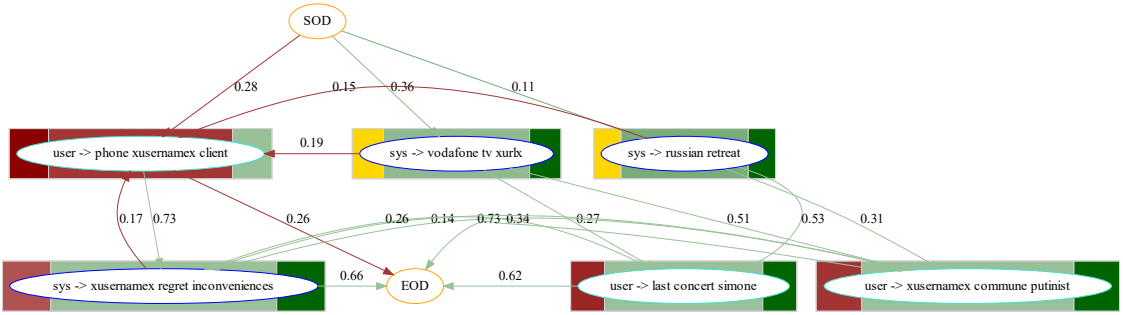

Figure 12: Sentiment-aware dialogue flow with standard deviation discovered for TDSAPT data

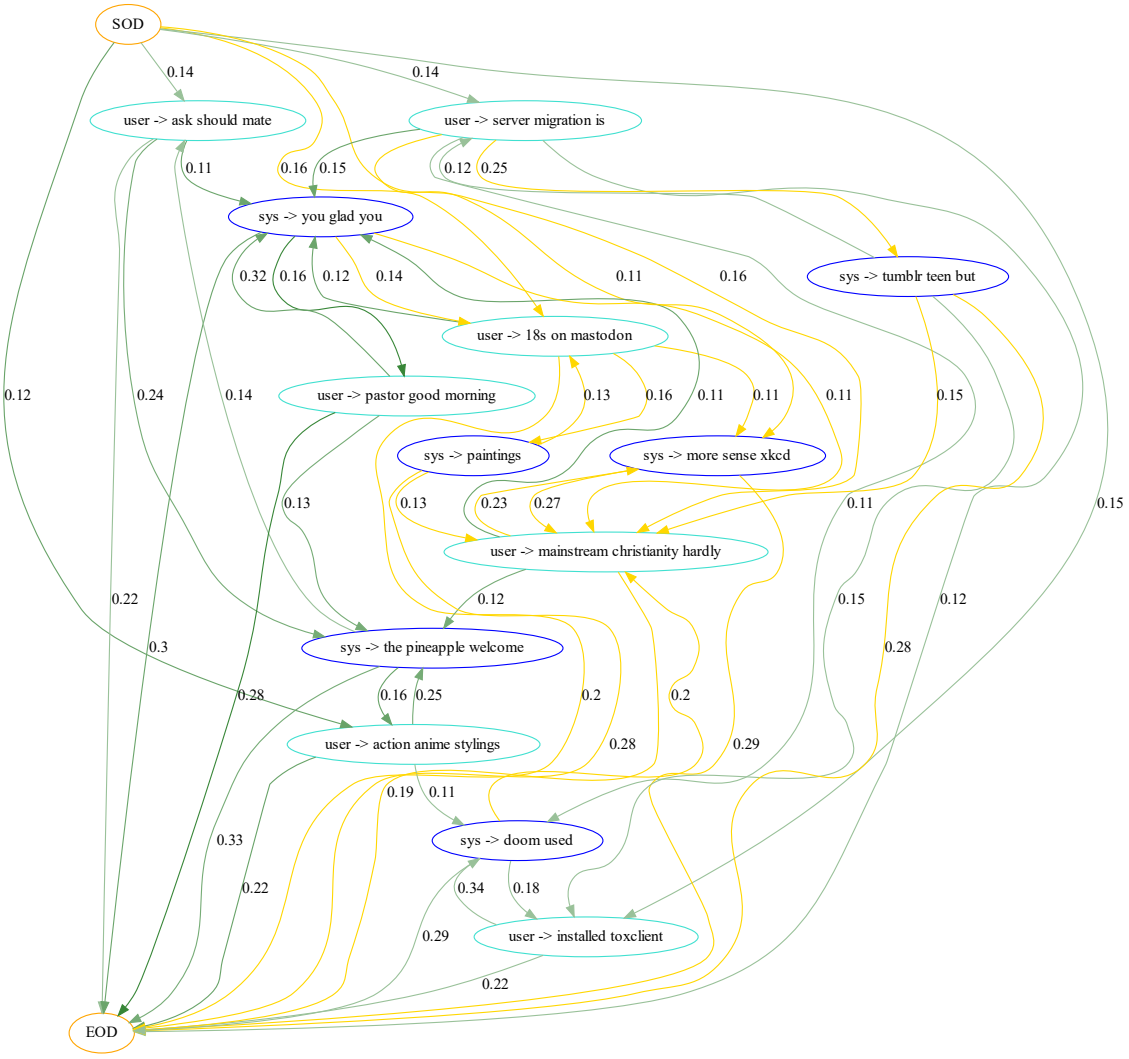

Figure 13: Sentiment-aware dialogue flow discovered for Mastodon data

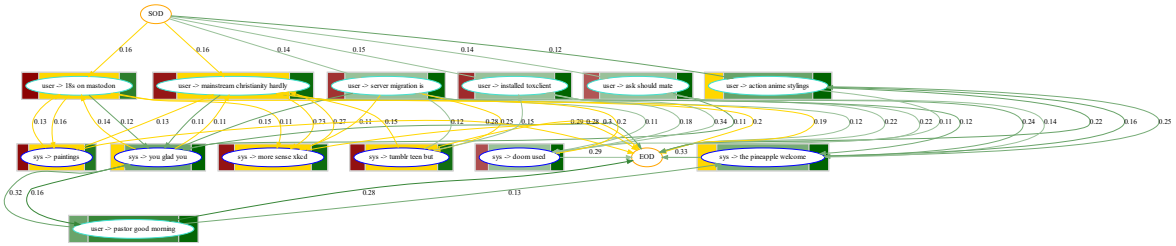

Figure 14: Sentiment-aware dialogue flow with standard deviation discovered for Mastodon data