Journal of Applied Economic Sciences
Volume XIX, Fall, Issue 3(85)

Enhancing Chatbot Intent Classification using Active Learning Pipeline for Optimized Data Preparation

Karolina KULIGOWSKA
https://orcid.org/0000-0001-8956-4723
Faculty of Economic Sciences, University of Warsaw, Poland
kkuligowska@wne.uw.edu.pl

Bartłomiej KOWALCZUK
https://orcid.org/0000-0003-2367-6268
Tidio Poland sp. z o.o., Szczecin, Poland
b.kowalczuk@tidio.net

Article's history:

Received 19th of July, 2024; *Received in revised form* 9th of August, 2024; *Accepted* 27th of August, 2024; *Available online*: 1st of September, 2024. *Published* as article in the Volume XIX, Fall, Issue 3(85), 2024.

Suggested citation:

Kuligowska, K., & Kowalczuk, B. (2024). Enhancing chatbot intent classification using active learning pipeline for optimized data preparation. *Journal of Applied Economic Sciences,* Volume XIX, Fall, 3(85), 317 – 325. https://doi.org/10.57017/jaes.v19.3(85).07

Abstract:

This study presents a novel approach to enhancing chatbot intent classification through an optimized data preparation combined with Active Learning. We applied the clustering mechanism using a state-of-the-art sentence-transformers model with cosine similarity for cluster detection in order to categorize messages. This process was further refined through a dedicated Active Learning Pipeline, which focused on the most essential observations for labeling. Incorporating externally sourced labeled data from Scale AI, the labeling process was fine-tuned iteratively, until the model's performance stabilized. This approach shows promise for various datasets and tasks, suggesting a scalable solution for preparing data for supervised modeling and achieving optimal model performance in real-world commercial chatbot scenarios.

Keywords: chatbot, intent classification, active learning, sentence-transformers, cluster label propagation.

JEL Classification: C45, C88.

Introduction

Active Learning has already a long history of research (Cohn et al., 1994; Lewis & Gale, 1994; Settles et al., 2008; Druck et al., 2009; Settles, 2009, 2011), similarly to creation and development of high-quality annotated corpora in the field of NLP (Dligach & Palmer, 2011; Settles, 2011; Grouin et al., 2014; Wissler et al., 2014). The quality of annotated data is essential for supervised learning paradigm in user intent detection and prediction for conversational artificial intelligence (Xiao et al., 2020; Pawlik et al., 2022; Chandrakala et al., 2024). Yet, there is still a lack of research covering commercial chatbots that include recent advances combining Active Learning and optimized data preparation for chatbot intent classification.

Our approach to enhance chatbot intent classification begins with a set of labeled data, then determines the adequate number of messages for the purpose of intent classification, and subsequently executes clustering of all first long messages of visitors represented as transformer-based embeddings, enabling further creation of main label categories. Next, the cluster label propagation was implemented for improving the labeling coverage within the dataset. Then, a dedicated Active Learning Pipeline was proposed to filter, refine, and reduce the dataset to the most informative observations. Additional labeled data from Scale AI were also incorporated, and the labeling process continued until incremental improvements in model performance began to flatten out with new data batches.

The rationale for our work arises from chatbots implemented for commercial use. The chatbot user's inquiries and requests regarding the products and services may cover hundreds of categories. Therefore, the chatbot intent classification model should quickly achieve the most accurate recognition of the intentions expressed in statements. Presented approach is trained considering state-of-the-art sentence-transformers model, cluster label propagation and Active Learning refinement. It is adaptable to various datasets and classification tasks, and has the ability to refine and prepare data for further supervised modeling in real-world commercial contexts.

The rest of this paper is organized as follows. Section 1 presents the relevant work in related research areas. Section 2 describes our approach to dataset creation for intent classification, while section 3 explains the proposed active learning pipeline. In last section we summarize our findings, and conclusions are included in the last section.

1. Related Literature Review

Our paper is related to research in three areas listed below.

Dataset creation for conversational agents

Creating datasets for building and developing conversational AI usually involves collecting text data from conversations and labeling users' messages in terms of the category resulting from their content (Hakkani-Tür et al., 2015). In case of proof of concept and very simple chatbot models, the categories may concern greeting recognition, inquiry or conversation termination (Setiaji & Wibowo, 2016). In case of chatbots implemented for commercial use, the categories of user messages may include up to several hundred detailed categories (Shafi et al., 2020; Vasquez-Correa et al., 2021) relating to the user's inquiries and requests regarding the products and services offered.

Intent classification in commercial chatbots

The labeled dataset serves as input for training and testing classification models. Due to the processing of unstructured text data, these models are rooted in natural language processing and machine learning. These include the use of intent classification models to understand in detail what the user wanted to achieve as a result of his/her message, and thus to achieve the most accurate recognition of the ideas expressed by users in their statements (Ouyang et al., 2022; Finch et al., 2023). For example, in messages such as: „what is the temperature outside this evening", „I need a ticket to Paris" or „the smartwatch I purchased has stopped working", the classification of intentions may consist of determining whether the user wants to get information about the weather, buy a ticket, or ask questions about the complaint process – and defining their appropriate categorization.

The ideal dataset created for intent classification for a commercial chatbot should contain a variety of user messages and be representative of the ways in which users may formulate their queries. According to Xiao et al. (2020), the more a chatbot is trained on the largest possible number of types of messages containing diverse intents that are present in text data extracted from conversations, the better it will understand the specific user intents and the better it will respond to a wide range of queries and requests for categories related to the dataset. Also, according to Chandrakala et al. (2024), having a large and comprehensive training database on which the chatbot's language processing model is trained allows it to more accurately handle more diverse and specific user queries. Therefore, increasingly advanced approaches are being developed and tested in chatbots (Suryanto et al., 2023) that are supposed to be efficient in interpreting customer interactions and understanding context.

Active learning

However, for any supervised machine learning model to perform well, it needs to be trained on thousands of labeled data. Unfortunately, very large sets of labeled text data are extremely time-consuming to produce and very expensive to obtain (Settles et al., 2008). Hence, attempts have been made to overcome this obstacle by developing new techniques involving machine learning to efficiently generate large amounts of labeled data in order to improve classification accuracy. One such method is active learning, also known in the statistical literature as „query learning" and „optimal experimental design" (Settles, 2009).

Active learning is a technique in which a model is actively involved in the process of acquiring a dataset (Druck et al., 2009). In practice, in the first step the model identifies the most informative observations and those that are difficult to classify. Then, in the second step, the most difficult cases are carefully selected, manually labeled by a human and added to the training set. By doing so, according to Settles et al. (2008) and Dligach & Palmer (2011), the model focuses on examples that are most likely to provide the greatest benefit when labeled and increase the accuracy of the model.

In addition to ensuring a balanced representation of category classes for the training and testing phases, it also aims to reduce the cost of manual data classification by limiting the procedure only to complex observations.

2. Dataset Creation for Intent Classification

Initial exploratory data analysis

The starting phase of the research involved an exploratory data analysis. For this purpose we used a dataset of 689,832 unique messages that we collected from December 2020 to January 2021. This dataset served as the initial corpus for extracting patterns and insights. A critical task of the starting phase, in order to ensure a representative subset of the data, was determining the adequate number of messages for the purpose of intent classification. We realized it during three subsequent iterations. They are summarized in Table 1 and described below.

Table 1. Initial EDA approach - selecting messages for intent classification

Iteration	Approach	Verification: result obtained	Improvement: strategy applied
1	Pairs of visitor-bot messages	Excessive noise in messages	Omit
2	First message of the visitor	Too short messages	Omit
3	First long message of the visitor: at least two words separated by whitespaces	Many-word message	Apply

Source: own elaboration.

First iteration was focused on analyzing pairs of visitor-bot interactions. However, this method proved suboptimal as the data contained excessive noise that complicated the extraction of meaningful patterns. Second iteration concentrated on the first message written by visitors. Analysis of this subset revealed that first messages, in most cases, were brief, often formulated as single-word messages, consisting of greetings such as „Hi" or „Hello". This method also proved suboptimal, as the dominance of short messages did not provide substantial analytical value. In response to the limitations observed in methods applied for the first two iterations, the third iteration set focus on the first long message of the visitor (at least two words separated by whitespaces). It proved optimal at this point, as covering a many-word message gives clearer insights. In fact, the first long message is the most descriptive one from the entire conversation. Hence, in most cases, it carries the most insightful information and is sufficient to assess the topic of the whole conversation. There are, of course, cases where the topic changes within a single conversation in the user's subsequent messages. Yet in practice, one conversation typically concerns one specific topic that reflects one clear intent of the user - which proves to be an extremely important observation from the business perspective. Moreover, applying this approach excluded brief greeting messages, which were identified as sources of noise and did not contribute to the analytical objectives.

After choosing the appropriate number of messages to classify intent, in order to improve the dataset and extract meaningful features from it, techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) and n-grams analysis were further employed. We realized it during two subsequent iterations. They are summarized in Table 2 and described below.

Table 2. Initial EDA approach - feature extraction for intent classification

Iteration	Technique	Verification: result obtained	Improvement: strategy applied
1	TF-IDF	Identification of redundant entities	Omit
2	N-grams analysis	Identification of redundant entities	Omit

Source: own elaboration.

While TF-IDF was utilized to assess the importance of words within the corpus, the n-grams analysis was utilized to capture the context by examining neighboring sequences of terms. Yet, the application of TF-IDF to the entire corpus failed to provide insightful results. It highlighted entities such as country names or product names, which did not provide deeper understanding of the text data. The subsequent n-grams analysis, focusing on the most frequent pairs of words, offered a marginally more informative perspective – but still it was too noisy and focused on redundant entities already highlighted by TF-IDF.

Clustering

Subsequently, we executed clustering of all first long messages of visitors. To achieve this, first we wanted to obtain vector representations of these first substantial messages through a language model. Due to their state-of-the-art performance, transformer-based language models are commonly used for this purpose in NLP (Ein-Dor et al., 2020; Wolf et al., 2020; Steegh & Sileno, 2023). After experimenting with several transformer models based on RoBERTa and DistilBERT, the sentence-transformers model elaborated by Reimers & Gurevych (2019), fine-tuned on quora dataset (Huggingface, 2019), proved to perform best on our data, assuming that we apply cosine similarity for cluster detection. Once we obtained vector representations, we executed clustering to group messages that share the same topics or contexts. We realized it during two subsequent iterations. They are summarized in Table 3 and described below.

Table 3. Clustering vector representations of visitors' first long messages

Iteration	Technique	Verification: result obtained	Improvement: strategy applied
1	UMAP, HDBSCAN, class-based TF-IDF	Generic clusters difficult to interpret, bias of redundant entities	Omit
2	Cosine-similarity between all message pairs	Small clusters, easy to interpret, minimized bias of redundant entities	Apply

Source: own elaboration.

First iteration was focused on a combination of advanced techniques, including Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) for clustering. This was followed by a class-based TF-IDF technique to highlight distinctive terms: all messages within a cluster were treated as a single document and then TF-IDF was calculated. Despite these sophisticated methods, the results were quite poor. The clusters that emerged were overly generic and difficult to interpret, and the bias of redundant entities was not minimized. Moreover, the whole iteration was compute-intensive. This outcome indicated a need for revising the approach to improve interpretability and efficiency.

In the second iteration of clustering vector representations of visitors' first long messages, we applied cosine similarity to assess the relationship between all pairs of messages. The clustering resulted in smaller groups, typically with up to 15 messages each, but most commonly around 5 messages per cluster. Small clusters were easy to interpret and included approximately 10,000 messages in total. This specific number of 10,000 represents the messages that were successfully clustered based on their relevance. Remaining messages were found to lack the necessary potential that would warrant their further inclusion by clustering algorithm.

This second iteration successfully minimized the bias of redundant entities, enhancing the quality and relevance of the data. Moreover, the second iteration was demanding on both memory and computational resources, requiring about 125 GB of RAM to process approximately 100,000 messages in a single pass. This high resource usage was primarily because we had to store all message embeddings simultaneously, therefore applying such a brute-force and not optimized approach resulted in the significant memory and computational requirements.

Manual labeling

Based on clusters of vector representations, the further creation of main categories of labels for the dataset was derived from:

- comprehensive analysis of past messages,
- common sense reasoning.

This approach led to the development of the first version of our taxonomy of labels, providing a foundation for further tasks. Manually assigning labels to each cluster resulted in 8,976 uniquely labeled messages, where each label could be applied on more than one cluster.

Cluster label propagation

To improve the labeling coverage within the dataset, the cluster label propagation was implemented with the idea of using existing labeled messages as references. For each message in the initial dataset, the cosine similarity was computed against all previously labeled messages, thus quantifying their overall similarity. If a message without a label revealed a high similarity to an already labeled message, this label was then assigned to the unlabeled message. Repeating this operation several times ensured that labels were systematically given to new messages that fulfilled semantic proximity. In result, the labeled dataset increased to 26,155 uniquely labeled messages.

Moreover, most probably due to similar nature of quora questions dataset of Reimers and Gurevych (huggingface, 2019) to our dataset, this model continuously exhibited really good results both on train and test sets, achieving 0.897 accuracy and 0.752 macro averaged f1 score on test set (randomly sampled 15% of the initial dataset, stratified by labels). Despite the high results of the model, it should be noted that these results were somewhat distorted because they were based on simple cases of „easy examples”. This issue was quickly exposed by an analysis conducted on a new sample that came from actual business operations. It turned out that the model's performance, which seemed good initially, was not as effective when applied to more complex real-world scenarios.

Therefore, we did not stop on cluster label propagation, and we decided to build a dedicated pipeline for the purpose of active learning to exclude easily classified messages and focus on more complex examples that could meaningfully contribute to model performance.

3. Active Learning Pipeline

The development of a model integrated with active learning was realized through a series of steps. As organization of the learning tasks in a way that they gradually progress to more complex ones is grounded in the literature of never-ending learning paradigm (Mitchell et al., 2018), we were focusing on optimizing targeted data utilization and iterative improvement.

In the first step a new dataset, aimed to provide a fresh basis for refining the model, was assembled. It included 225,323 unique English messages sourced from distinct conversations originating in June, 2021. It is worth underlining that the name of the month is quite important here, because including messages from summer could slightly reduce the impact of seasonality. The previous sample originated from December and January, where messages were heavily concentrated around holidays. Analyzing this thread in a broader perspective could contribute to research in another direction, concerning the cultural environment where the chatbot operates, and focusing on its impact on seasonality in training and testing data.

Then, to maximize the impact of the data on model performance, a dedicated Active Learning Pipeline was proposed. The pipeline developed by us was designed to filter, refine, and prepare data for further modeling; subsequently – to reduce the dataset to the most informative samples; and finally - to ensure a balanced representation of classes for training and evaluation phases. The assumed strategy was to select messages that were most likely to improve learning and enhance model accuracy:

- The pipeline started with the detection of the language of incoming messages. We decided to use an open-source tool such as FastText. Given the project's focus, only messages identified as English were retained. After comparing several algorithms of language detection (including FastText, LSTM_langid, langdetect, and Google's cld3), FastText turned out to be the best available option. Nevertheless, it is worth adding that it remains a weak point in the project due to limitations in accurately identifying language nuances.
- In the first few iterations, predictions were computed using the sentence-transformers model fine-tuned on quora dataset – the same model that was improved by the cluster label propagation step.
- Then to our dataset, we applied heuristics based on regular expressions. These heuristics were created based mainly on essential keywords that could include specific predicted classes (e.g. for order status, the keyword is „tracking number” etc.). These heuristics covered 70% of the dataset. Messages that could be classified through these heuristics were removed from further analysis - under the assumption that they consist of „easy examples” from which the model would derive minimal learning benefit.
- Next, remaining messages were further refined by applying a threshold to the model's scores. The assumed strategy was to reduce the remaining sample by thresholding on the model's outputs. According to our findings, messages with a maximum class score below the threshold of 0.8 should be considered as „difficult examples” due to the model's uncertainty, indicated by a flatter distribution of scores. These complex „difficult” messages were flagged for manual labeling, a decision supported by subject literature (Zhang et al., 2023; Meer et al., 2024) and for own manual analysis that distinguished a clear density difference in examples above and below this threshold.
- Finally, from the refined messages, we drew stratified samples. With fewer labeled data batches, we expected them, after Settles (2011), to be more economical and more efficient at achieving high accuracy. Based on the predicted classes we ensured a representative composition in each batch. The first few samples included 1000 observations each.

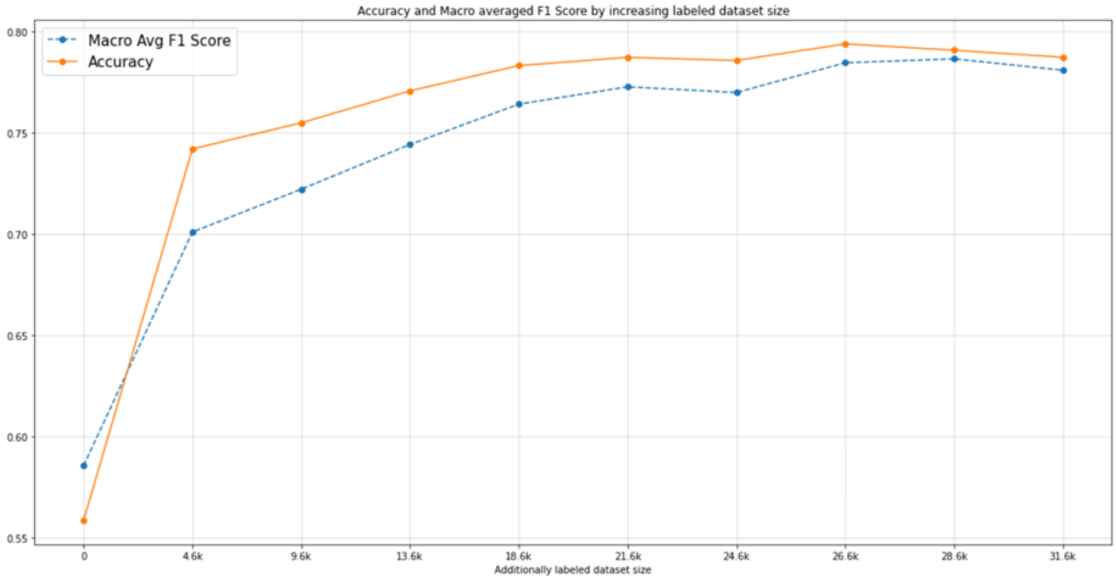
After realizing steps of Active Learning Pipeline, we wanted to assess the performance of the model. It is worth mentioning that each individual data sample needed to be annotated by Scale AI before proceeding to the next stage. Therefore, the active learning workflow could only advance once Scale AI had provided the necessary labels for each sample, ensuring that the dataset was properly annotated. Thus, having accumulated around 10,000 labeled messages sourced from Scale AI, we refined the taxonomy and validation set of 5,128 messages, stratified on labels. Then, we fine-tuned the model on the combined dataset, focusing on messages covered by heuristics for higher accuracy.

We iteratively fine-tuned the model, validated performance, monitored errors, and targeted weak points with specific labeled samples. Each iteration improved the model, leading to the replacement of the previous version. To optimize the pipeline and data quality, we removed messages from the „undefined" class with high confidence, lowered the threshold to capture more complex cases, and sampled missing or difficult categories. We also extended regex heuristics throughout the process.

The labeling process was concluded once the incremental improvements in model performance began to flatten out with the addition of new data batches. The incremental improvements slowed down or even decreased after adding new batches, as presented on Figure 1.

This step also revealed biases in the annotated data due to differences of view between taskers. Namely, as some categories are very similar, taskers assigned labels from their own perspective - which was not always the same as our perspective. In several cases it was leading to an erroneous process of inconsistent label assignments. Because of this, we decided that adding more data would not improve the model's performance, according to the earlier evidence discussed in research on human annotation efforts to build reliable annotated datasets (Settles, 2011; Grouin et al., 2014).

Figure 1. Graph presents crucial metrics of the model fine-tuned on increasing labeled dataset



Source: own elaboration.

The final step involved merging several product categories into a single category. We have observed a high incidence of errors within these categories, but not between them and the remaining categories. It was determined that treating these as one consolidated category would be the most effective for the current model iteration. This was supported not only by considerations from the perspective of machine learning engineering, but also by the sufficient business productivity of the entire model pipeline.

4. Research Results

Our research indicates that conventional text analysis methods, such as TF-IDF and n-grams analysis, may fall short when it comes to uncovering deeper, contextually significant insights within certain datasets. We found that clustering techniques offer substantial advantages by revealing patterns that are not immediately visible through simple text analysis. Additionally, cluster label propagation improves labeling consistency and scalability, although its effectiveness is closely tied to the quality and representativeness of the initially labeled messages. Any errors or biases in these original labels can potentially be propagated throughout the dataset, making the initial label quality a critical consideration.

To address these challenges, we developed the Active Learning Pipeline. This pipeline uses regex-based heuristics to filter out easily classified messages, thereby focusing on more complex examples that contribute meaningfully to model performance. By applying a threshold to model outputs, the pipeline emphasizes data points where the model exhibits uncertainty, which can lead to significant improvements in accuracy. Moreover, stratified sampling based on predicted classes ensures that all classes are adequately represented in the dataset.

However, the overall effectiveness of this pipeline strongly depends on the accuracy of the initial model predictions. Biases or inconsistencies in the initial model can be amplified in subsequent steps, potentially leading to suboptimal results. Additionally, while regex heuristics streamline the process, they might inaccurately discard potentially valuable messages. Similarly, a threshold set rigidly for model scores can exclude messages that are near the threshold but still hold significant potential.

Conclusion

Main contribution of this study is the development of the Active Learning Pipeline designed to enhance commercial chatbots performance by improving the dataset quality for classification models in user intent detection.

The pipeline started with the language identification of incoming messages. Then, the initial iterations use a preliminary model, which was subsequently refined through cluster label propagation. Heuristics, based on regular expressions and essential keywords indicative of specific classes, were then applied to approximately 70% of the dataset. The remaining messages, categorized as „difficult examples", were flagged for manual annotation. Subsequently, stratified sampling was performed on these refined messages to ensure a representative composition in each batch. Following the implementation of the Active Learning Pipeline, the taxonomy was revised, and the model was then fine-tuned on a combined dataset, including messages addressed by heuristics.

The results indicate that traditional text analysis methods (such as TF-IDF and n-grams) turn out to be insufficient for extracting contextually significant insights in commercial chatbot datasets. The Active Learning Pipeline, particularly with the integration of cluster label propagation, significantly improves dataset quality and classification performance. It reduces the manual effort required to distinguish between „easy" and „difficult" examples from extensive datasets, focusing manual labeling efforts on more challenging cases.

This approach offers practical value as it provides the baseline for methodologies for creating and refining datasets to improve the intent recognition, and enhancing active learning of the linguistic layer of commercial chatbots.

Credit Authorship Contribution Statement

Both authors play equal roles in the research and development of the project, including the analysis, writing, and decision-making processes. Their contributions were collaborative and evenly distributed throughout the study.

Acknowledgments/Funding

This work was co-financed by the National Centre for Research and Development (NCBR) POIR.01.01.01-00-0455/19 and presents partial results of a research project entitled „Building advanced chatbots (version 2.0) dedicated to automating corporate customer service" carried out by Tidio Poland sp. z o.o.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

[1] Chandrakala, C. B., Bhardwaj, R., & Pujari, C. (2024). An intent recognition pipeline for conversational AI. *International Journal of Information Technology*, 16, 731-743. https://doi.org/10.1007/s41870-023-01642-8

[2] Cohn, D., Atlas, L. & Ladner, R. (1994). Improving generalization with active learning. *Machine Learning*, 15(2), 201–221. https://doi.org/10.1007/BF00993277

[3] Dligach, D., & Palmer, M. (2011). Reducing the Need for Double Annotation. [in:] Ide, N., Meyers, A., Pradhan, S., & Tomanek, K. (eds.). *Proceedings of the 5th Linguistic Annotation Workshop (LAW-V).* ACL SIGANN, USA, 65–73. https://aclanthology.org/W11-0408

[4] Druck, G., Settles, B., & McCallum, A. (2009). Active learning by labeling features. [in:] *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP '09). ACL, Singapore, 81–90. https://dl.acm.org/doi/10.5555/1699510.1699522

[5] Ein-Dor, L., Halfon, A., Gera, A., Shnarch, E., Dankin, L., Choshen, L., Danilevsky, M., Aharonov, R., Katz, Y., & Slonim, N. (2020). Active Learning for BERT: An Empirical Study. [in:] Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). ACL, 7949-7962. https://doi.org/10.18653/v1/2020.emnlp-main.638

[6] Finch, S. E., Paek, E. S., & Choi, J. D. (2023). Leveraging Large Language Models for Automated Dialogue Analysis. [in:] S. Stoyanchev, S. Joty, D. Schlangen, O. Dusek, C. Kennington, & M. Alikhani (eds.), *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. ACL. 202-215. https://doi.org/10.18653/v1/2023.sigdial-1.20

[7] Grouin, C., Lavergne, T., & Névéol, A. (2014). Optimizing annotation efforts to build reliable annotated corpora for training statistical models. [in:] Levin L., Stede M. (eds.). Proceedings of the 8th Linguistic Annotation Workshop (LAW-VIII). ACL SIGANN, Ireland, 54–58. https://doi.org/10.3115/v1/W14-4907

[8] Hakkani-Tür, D., Ju, Y.-C., Zweig, G., & Tur, G. (2015). Clustering Novel Intents in a Conversational Interaction System with Semantic Parsing. [in:] Proceedings of Interspeech, Germany. 1854-1858. https://doi.org/10.21437/Interspeech.2015-70

[9] Huggingface (2019). https://huggingface.co/sentence-transformers/distilbert-base-nli-stsb-quora-ranking [accessed 08.2024]

[10] Lewis, D. D., Gale, W. A. (1994). A Sequential Algorithm for Training Text Classifiers. [in:] Croft, B.W., van Rijsbergen, C. J. (eds.), SIGIR '94, *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. Springer, London. 3-12. https://doi.org/10.1007/978-1-4471-2099-5_1

[11] Meer, M. van der, Falk, N., Murukannaiah, P. K., & Liscio, E. (2024). Annotator-Centric Active Learning for Subjective NLP Tasks. ArXiv, 1-18. https://doi.org/10.48550/arXiv.2404.15720

[12] Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Yang, B., Betteridge, J., Carlson, A., Dalvi, B., Gardner, M., Kisiel, B., Krishnamurthy, J., Lao, N., Mazaitis, K., Mohamed, T., Nakashole, N., Platanios, E., Ritter, A., Samadi, M., Settles, B., Wang, R., Wijaya, D., Gupta, A., Chen, X., Saparov, A., Greaves, M., & Welling, J. (2018). Never-ending learning. Communications of the ACM, 61(5), 103–115. https://doi.org/10.1145/3191513

[13] Nguyen, D. H. M., & Patrick, J. D. (2014). Supervised machine learning and active learning in classification of radiology reports. *Journal of the American Medical Informatics Association*, 21(5), 893-901. https://doi.org/10.1136/amiajnl-2013-002516

[14] Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C. L., Mishkin P., Zhang C., Agarwal S., Slama K., Ray A., Schulman J., Hilton J., Kelton F., Miller L., Simens M., Askell A., Welinder P., Christiano P. F., Leike J., & Lowe R. (2022). Training language models to follow instructions with human feedback. [in:] Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022), 1-15.

[15] Pawlik, Ł., Płaza, M., Deniziak, S., & Boksa, E. (2022). A method for improving bot effectiveness by recognising implicit customer intent in contact centre conversations. *Speech Communication*, 143, 33-45. https://doi.org/10.1016/j.specom.2022.07.003

[16] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. [in:] *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 1-11. https://doi.org/10.48550/arXiv.1908.10084

[17] Setiaji B., Wibowo F. W. (2016). Chatbot using a knowledge in database: Human-to-machine conversation modeling. [in:] *Proceedings of the 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS)*, IEEE, 72-77. https://doi.org/10.1109/ISMS.2016.53

[18] Settles, B. (2009). Active Learning Literature Survey. Computer Sciences Technical Report 1648. University of Wisconsin–Madison, 1-44.

[19] Settles, B. (2011). From Theories to Queries: Active Learning in Practice. [in:] Guyon, I., Cawley, G., Dror, G., Lemaire, V., & Statnikov, A. (eds.). Active Learning and Experimental Design Workshop in conjunction with AISTATS 2010, Proceedings of Machine Learning Research, 16, 1-18.

[20] Settles, B., Craven, M. W., & Friedland, L. A. (2008). Active Learning with Real Annotation Costs. [in:] Proceedings of the NIPS Workshop on Cost-Sensitive Learning, 1-10.

[21] Shafi, P. M., Jawalkar, G. S., Kadam, M. A., Ambawale, R. R., & Bankar, S. V. (2020). AI-Assisted Chatbot for E-Commerce to Address Selection of Products from Multiple Products. [in:] Dey, N., Mahalle, P., Shafi, P., Kimabahune, V., & Hassanien, A. (eds.). *Internet of Things, Smart Computing and Technology: A Roadmap Ahead. Studies in Systems, Decision and Control,* 266, 57-80. https://doi.org/10.1007/978-3-030-39047-1_3

[22] Steegh, E., & Sileno, G. (2023). No labels? No problem! Experiments with active learning strategies for multi-class classification in imbalanced low-resource settings. [in:] *Proceedings of the 19th International Conference on Artificial Intelligence and Law* (ICAIL '23). ACM, Portugal, 277-286. https://doi.org/10.1145/3594536.3595171

[23] Suryanto, T., Wibawa, A., Hariyono, H., & Nafalski, A. (2023). Evolving Conversations: A Review of Chatbots and Implications in Natural Language Processing for Cultural Heritage Ecosystems. *International Journal of Robotics and Control Systems,* 3(4), 955-1006. https://doi.org/10.31763/ijrcs.v3i4.1195

[24] Vasquez-Correa, J. C., Guerrero-Sierra, J. C., Pemberty-Tamayo, J. L., Jaramillo, J. E., & Tejada-Castro, A. F. (2021). One System to Rule Them All: A Universal Intent Recognition System for Customer Service Chatbots, 1-26. http://dx.doi.org/10.2139/ssrn.3986692

[25] Wissler, L., Almashraee, M., Monett Diaz, D., & Paschke, A. (2014). The Gold Standard in Corpus Annotation. [in:] Proceedings of the 5th IEEE Student Conference, IEEE, Germany, 1-4. https://doi.org/10.13140/2.1.4316.3523

[26] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. [in:] Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP). ACL, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[27] Xiao, Z., Zhou, M. X., Chen, W., Yang, H., & Chi, C. (2020). If I Hear You Correctly: Building and Evaluating Interview Chatbots with Active Listening Skills. [in:] *Proceedings of the CHI Conference on Human Factors in Computing Systems* (CHI '20). ACM USA. 1-14. https://doi.org/10.1145/3313831.3376131

[28] Zhang, Z., Strubell, E., & Hovy, E. H. (2022). A Survey of Active Learning for Natural Language Processing. ArXiv. 1-26. https://doi.org/10.48550/arXiv.2210.10109