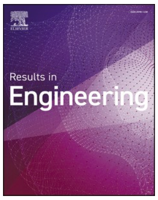
Results in Engineering 17 (2023) 100856

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering



A comparison of chatbot platforms with the state-of-the-art sentence BERT for answering online student FAQs

Kevin Peyton [a,*], Saritha Unnikrishnan [a,b]

[a] Faculty of Engineering and Design, Atlantic Technological University, Sligo, Ireland
[b] Mathematical Modelling and Intelligent Systems for Health and Environment (MISHE), Atlantic Technological University, Sligo, Ireland

ARTICLE INFO

Keywords:
Chatbots
SBERT
Natural language understanding
FAQs
Online learning

ABSTRACT

Online learning enables academic institutions to accommodate increased student numbers at scale. With this scale comes high demands on support staff for help in dealing with general questions relating to qualifications and registration. Chatbots that implement Frequently Asked Questions (FAQs) can be a valuable part in this support process. A chatbot can provide constant availability in answering common questions, allowing support staff to engage on higher value one-to-one communication with prospective students. A variety of approaches can be used to create these chatbots including vertical platforms, frameworks, and direct model implementation. A comparative analysis is required to establish which approach provides the most accuracy for an existing, available dataset.

This paper compares intent classification results of two popular chatbot frameworks to a state-of-the-art Sentence BERT (SBERT) model that can be used to build a robust chatbot. A methodology is outlined which includes the preparation of a university FAQ dataset into a chatbot friendly format for upload and training of each implementation. Results obtained from the framework-based implementations are generated using their published Application Programming Interfaces (APIs). This enables intent classification using testing phrases and finally comparison of F1 scores.

Using ten intents comprising 284 training phrases and 85 testing phrases it was found that a SBERT model outperformed all others with an F1-score of 0.99. Initial comparison with the literature suggests that the F1-scores obtained for Google Dialogflow (0.96) and Microsoft QnA Maker (0.95) are very similar to other benchmarking exercises where NLU (Natural Language Understanding) has been compared.

1. Introduction

Chatbots are intelligent conversational agents, which have become an essential element of the customer management process for dealing with large numbers of online queries and support tasks. Customer service has evolved to the point where traditional voice and email interactions are now being replaced by self-service channels such as chatbots. According to van der Goot and Pilgrim [1] such channels are acceptable to customers once queries are answered in a fast, easy and convenient manner.

While chatbots are being widely implemented in sectors such as telecoms, finance and online commerce, Yang and Evans [2] note the absence of implementation in the educational sector. With online learning becoming ever more popular, it seems that the ‘24/7’ availability highlighted by Cunningham-Nelson et al. [3] can be extremely advantageous. By its very nature, online learning does not necessarily confine a student to synchronous engagement or through shared geography or time zone. However, it does place considerable burden on the academic institutions as to how they scale their support to prospective students with questions relating to qualifications, fees, and other administrative queries.

Using chatbots to provide a dependable solution for answering common questions and queries, allows support staff to be better utilised in interacting with prospective students on a personal, one to one level. There are many approaches that can be used to support the delivery of a chatbot that can deliver FAQs, and these include vertical platforms

[*] Corresponding author.
E-mail addresses: kevin.peyton@atu.ie (K. Peyton), saritha.unnikrishnan@atu.ie (S. Unnikrishnan).

https://doi.org/10.1016/j.rineng.2022.100856
Received 27 October 2022; Received in revised form 13 December 2022; Accepted 17 December 2022
Available online 21 December 2022

relevant to the education sector such as Mainstay[1] and more well-known platforms such as Dialogflow[2] from Google and LUIS[3] from Microsoft. Direct implementation of a state-of-the-art (SOTA) technique for Natural Language Processing (NLP) may also be an option. While these approaches may differ in terms of market segment, technical approach, and user environment – they all have a common factor which is how effective and accurate the results are regarding Natural Language Understanding (NLU).

This paper provides a comparative analysis of popular chatbot platforms with a SOTA SBERT implementation to establish which approach provides the highest NLU accuracy for an FAQ dataset associated with queries from prospective students applying to study online. Under Related work, approaches are outlined in the literature to NLU accuracy that have already been carried out. These appear to be mainly from generic, open domains such as travel – for example using station names along with arrival and departure times which potentially enable easier intent identification. In this work, the dataset being evaluated will be specific to a subset of queries within an educational domain for prospective students to study online. A comparative approach using this type of dataset has not been found in the literature. In the Methodology section, the frameworks and the models used are briefly described. A detailed account is also given of dataset creation and implementation of a test harness for generating results. In the Results & Discussion section, results are discussed and compared to studies from the literature.

2. Related work

A review of publications that discuss benchmarking and analyse the results of NLU accuracy suggest several approaches. The comparison of results and features from platforms like Watson[4] from IBM, Dialogflow, LUIS, and RASA[5] is common and suggests that general question answering services are well served by these platforms. It was found that the datasets being compared are primarily from open domains utilising relatively common intents such as those for restaurant booking and travel reservations.

This approach in using open domains was illustrated by Braun et al. [4] when they investigated LUIS and used queries from travel (206 queries) and online computing forums (290 queries). Wisniewski et al. [5] compared several systems including LUIS for building chatbots, using an open domain dataset comprising 328 queries. By contrast Liu et al. [6] performed a cross domain implementation of NLU using 21 domains comprising 25 k queries whilst noting the difficulty the user might have in choosing between platforms.

A different approach taken by Malamas et al. [7] who used the RASA platform within a healthcare domain and modified model internal settings. This experimentation used a hand designed dataset of 142 questions with the author commenting that the collection, review and adding of new data is quite normal within the domain. Intriguingly, they also noted that intent similarity can be an issue where a sentence may only differ by one or two words.

3. Methodology

3.1. Frameworks and models used

This study compared the NLU results of chatbot platforms from Microsoft and Google to those from a traditional Feedforward model, and a SOTA Sentence BERT[6] (SBERT) model. QnA Maker[7] from Microsoft and Dialogflow from Google, are both cloud-based NLP services that allows for the creation of conversational client applications, including chatbots. These are suitable for use when static, non-changing information like FAQs are being used to return answers to the user. They both enable the importation of structured content (in the form of question and answer pairs) as well as semi-structured content (FAQs, manuals, documents) to a knowledge base. It is through this knowledge base that the user can easily manipulate and improve the information by adding different forms of how the question might be asked as well as metadata tags that are associated with each question and answer pair.

Both QnA Maker and Dialogflow provide rich environments for authoring, building and publishing conversational clients by both developers and non-developers. In this study, these environments were used for the upload and training of our data on both platforms. In the case of QnA Maker, NLU is provided by direct integration into LUIS. Along with QnA Maker and Dialogflow, a traditional Feedforward neural network model was utilised. Fig. 1 shows an implementation of one of these networks with one input, eight hidden layers and one output.

The implementation in this study is based on the work by Loeber [8]. It used Python 3.7 with PyTorch and nltk libraries and utilised hyperparameters including batch size of 8, learning rate of 0.001, hidden size of 8 and 1000 epochs for training. Finally, an SBERT pre-trained transformer framework developed by Reimers and Gurevch [9] was implemented to evaluate the dataset in this study. Fig. 2 shows an example of the SBERT architecture computing a similarity score between two sentences to see how similar in meaning the sentences are.

At the core of the SBERT model is the pretrained BERT network developed by Devlin et al. [11] that enables bidirectional training to be applied to a word sequence, using a transformer. A transformer uses an encoder-decoder architecture with self-attention, a mechanism that enables the model to make sense of the input that it receives. This ability to make sense of language sequences has allowed BERT to be used for different types of tasks including classification, regression, and sentence similarity. However, it performs poorly with semantic similarity search, a task that can be used to measure the similarity score of texts and sentences in terms of a defined metric. Whereas the construction of BERT makes it unsuitable for semantic similarity search, Reimers and Gurevych [9] demonstrated that SBERT could perform this more efficiently by deriving semantically meaningful sentence embeddings or

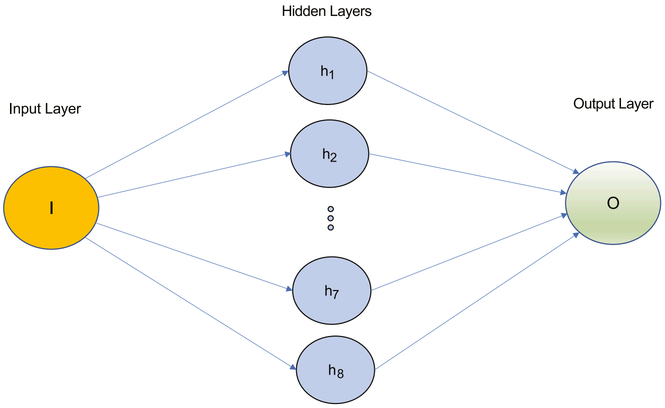


Fig. 1. A Feedforward neural network with one input, eight hidden layers and one output.

[1] https://mainstay.com/.
[2] https://cloud.google.com/dialogflow/.
[3] https://www.luis.ai/.
[4] https://www.ibm.com/cloud/ai.
[5] https://rasa.com/.
[6] https://sbert.net.
[7] https://www.qnamaker.ai/.

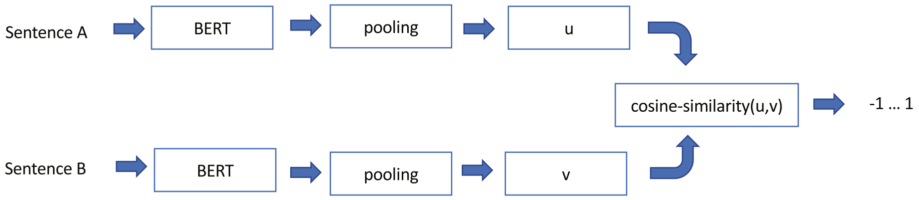


Fig. 2. SBERT Architecture showing Sentence A and Sentence B passed through the network yielding embedding U, V with cosine similarity then applied [10].

vectors, that can be compared. Cosine similarity is used for this comparison process as it enables a measure of similarity to be ascertained about documents or words in a document. It offers the potential for improved text retrieval by understanding the content of the supplied text. This model is particularly interesting to the current study as a human may pose a question in a variety of ways. While the words may be different, they may be semantically similar in terms of meaning.

In this implementation of SBERT, Python 3.7. was used, with the sentence-transformer library framework utilising the “bert-base-nli-mean-tokens” pre-trained model. Training of the models and all testing for the platforms and models outlined here was conducted on a MacBook Intel Core i5 with 8 GB RAM.

3.2. Dataset

The data used for this study originates in question and answers format from the online FAQs that are available for prospective online students at a Technological University. Fig. 3 shows an example of how this material was presented to the user on the website in a collapsible format that allowed easy access of answers to common questions.

This material was originally written to be quickly scanned by the student for answers to commonly asked queries received by administrative staff via phone and email. The dataset was divided into three categories, entitled About, Applications and Studying Online. A fourth category encompassing a special purpose Government funded training initiative known as Springboard+ [12] was also included. These four categories comprised a total of 41 question answer pairs. Table 1 outlines the original publishing details of the question answer pairs showing an overall minimum question length of three words and an overall maximum question length of sixteen words.

The existing structure and content of the data was perfect for easy assimilation by a human scanning for relevant information. It was recognised that this format was not optimal for chatbot usage. For example, some answers were quite long, and others contained information not directly relevant to the question being asked from the perspective of a potential chatbot response.

For this reason, the initial content was used as the basis for a revised set of question answer pairs, optimised for use in a generic storage format. Fig. 4 shows a revised, generic question answer pair format utilising JavaScript Object Notation (JSON) that allows for storage of potential questions that will be asked. Each question or utterance and associated data is stored in a node using a key/value, individual value or array of values [13].

Table 2 describes the specific structure and data related to an utterance. This structure allows multiple utterances in the form of an FAQ to be easily integrated and tested with disparate chatbot frameworks and models.

An individual utterance should have a unique tag, but a category will normally contain multiple, related utterances. A test utterance is never used to train the chatbot – it is used to only to test the effectiveness of a chatbot trained on utterances provided as patterns. Finally, depending on how the chatbot operates – a response is generated for the user and this response will have an associated confidence score.

Table 3 above shows the reformatted and rewritten question answer pairs used for chatbot purposes. The total number of training phrases or questions is 294. This allows better tailored and more specific responses to potential user questions and also ensures that more tailored training data is available. The total number of question answer pairs has grown from the original 41 to a total of 85. The question part of these 85 pairs are the test questions used in the testing of the models.

3.3. Implementation

A highly focused and complete dataset, as described in Section 3.2, was used to evaluate the NLU performance on the Dialogflow and QnA Maker platforms and the Feedforward and SBERT models. Fig. 5 outlines the details of the simple test harness that was built to utilise common processes that could be applied across all four approaches.

The key processes included importing of the training data, evaluating test data, and generating a standard output of results. The storage of the training data was done by using a standard JSON format as discussed in section 3.2. Similarly, in extracting data from this training set to run the test data through the implementations, a bespoke Python library was developed to read the dataset and generate output results.

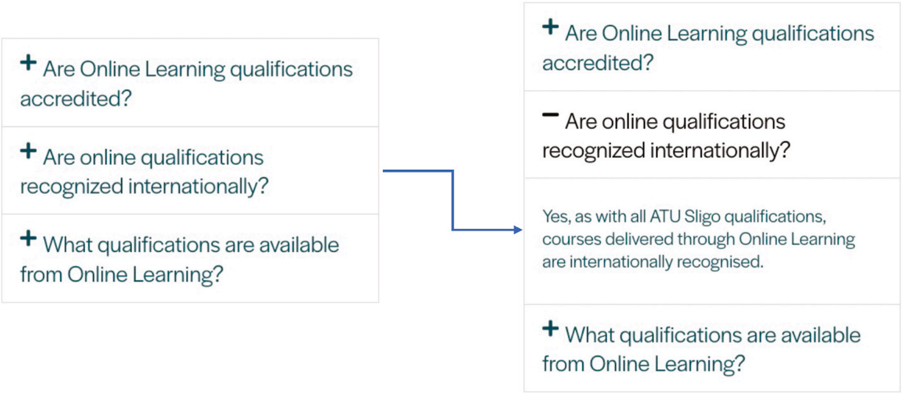
With the import of training data – both Dialogflow and QnA Maker offered straightforward import options for training data. Python scripts were written to transform the JSON dataset into CSV (comma separated) and TSV (tab separated) files respectively with the training process on both platforms being initiated manually. As both the Feedforward and SBERT implementations were bespoke, full control of the automation and associated scripting was possible, with import of the JSON data and training of the model occurring in one manual cycle. Whereas the former was straightforward in terms of implementation, the latter was more complex. A pre-trained SBERT model (bert-base-nli-mean-tokens) was implemented for semantic search. To store the subsequent trained embeddings and to enable future semantic search an Elasticsearch[8] container running version 7.9.0 was provisioned using Docker.[9]

Running test data through the implementations followed a similar convention as described in the training process. Both Dialogflow and QnA Maker offer well documented APIs which was scripted using a bespoke Python library to query the trained models with test questions. Feedforward and SBERT models were evaluated in a similar fashion. Finally, the results from all implementations outputted data in a standard format. This format was again generated across all implementations using a combination of Python scripts and Libraries to generate generic logging data, F1 scores and the graphical generation of a confusion matrix. The generic logging data included question asked, answer expected, answer predicted, confidence, intent expected, and intent predicted. A sample of this logging data is shown in Fig. 6.

4. Results and discussion

Table 4 shows a comparison of the evaluated model results with relevant Precision, Recall and F1-scores. Looking at these results in isolation it appears that SBERT gives the most promising results with an F1 score of 0.99 on this particular dataset. This was followed closely by Dialogflow with an F1 score of 0.96 and QnA Maker with an F1 score of 0.95. Feed Forward comes in unsurprisingly at the end with an F1 score

[8] https://www.elastic.co/.
[9] https://www.docker.com/.



+ Are Online Learning qualifications accredited?

+ Are online qualifications recognized internationally?

+ What qualifications are available from Online Learning?

+ Are Online Learning qualifications accredited?

− Are online qualifications recognized internationally?

Yes, as with all ATU Sligo qualifications, courses delivered through Online Learning are internationally recognised.

+ What qualifications are available from Online Learning?

Fig. 3. Original format of FAQs on website used accordion for collapsible content.

Table 1
Question answer pairs of original website text.

Section	Question Answer pairs	Question min length	Question max length
About	3	5	7
Applications	8	5	12
Studying Online	14	5	15
Springboard+	15	3	16
TOTAL	**41**		

of 0.56.

SBERT, Dialogflow and QnA Maker appear to have similar issues contained within the sb_intro and sb_apply intents of this dataset. Each of these intents contain seventeen and nine intents respectively and have a slightly lower F1 score when compared with other intents across the implementations. This may relate to an observation noted by Malamas et al. [7] who stated that intent similarity can be an issue. In this particular case, it is the fact that the phrase "springboard" must be used in every question relating to these specific intents.

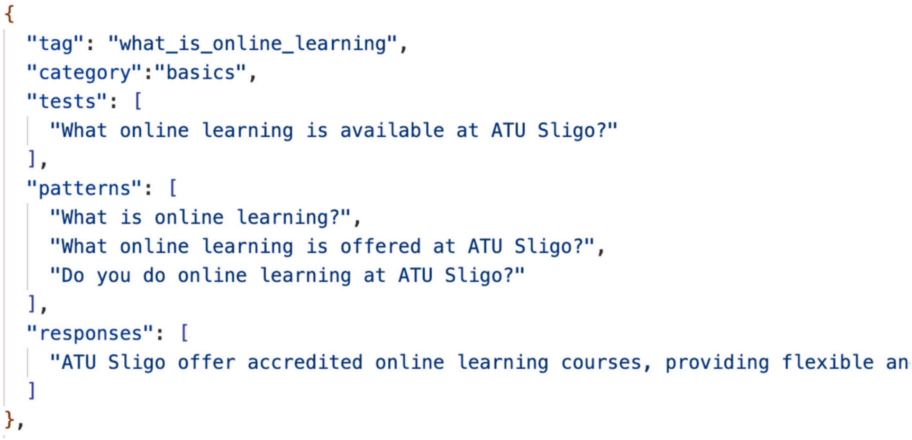
While the results outlined indicate that SBERT with this particular dataset perform well, it is worthy to consider the performance of Dialogflow and QnA Maker in the literature. Braun et al. [4] considered LUIS as part of their evaluation and found it to give their highest overall F1 score of 0.916 between compared platforms. Liu et al. [6] considered both LUIS and Dialogflow as part of their benchmarking and reported overall F1 scores of 0.821 and 0.811 respectively. Broad comparisons of our F1 scores with those in the literature are however less relevant for several reasons. Firstly, online platforms and underlying models may

Table 2
Specific structure and data related to an utterance.

Name	Description
Tag	Unique name for the utterance
Category	A category may contain multiple utterances and equates to intent
Tests	Utterance(s) provided to test the effectiveness/accuracy of a chatbot.
Patterns	Utterance used to train the chatbot. A pattern shows that an utterance can be phrased in several different ways, comparable to how a real use might ask a question
Responses	Each utterance normally contains just one response. A salutation utterance might randomly choose from a number of responses (e.g., pattern: 'Hello', response: "Hi" or "Hi there")

Table 3
Reformatted and rewritten question answer pairs.

Category/Intent	Question Answer pairs	Original Section
Basics	9	About/Studying Online
Dates	5	About
Applications	12	Applications
Study	10	Studying Online
exam_ca	6	About/Studying Online
Fees	6	About
sb_intro	17	Springboard+
sb_apply	9	Springboard+
sb_fees	6	Springboard+
sb_details	5	Springboard+
TOTAL	**85**	**294 training phrases**



```
{
  "tag": "what_is_online_learning",
  "category":"basics",
  "tests": [
    "What online learning is available at ATU Sligo?"
  ],
  "patterns": [
    "What is online learning?",
    "What online learning is offered at ATU Sligo?",
    "Do you do online learning at ATU Sligo?"
  ],
  "responses": [
    "ATU Sligo offer accredited online learning courses, providing flexible an
  ]
},
```

Fig. 4. JSON snippet showing node structure of sample question answer pair.

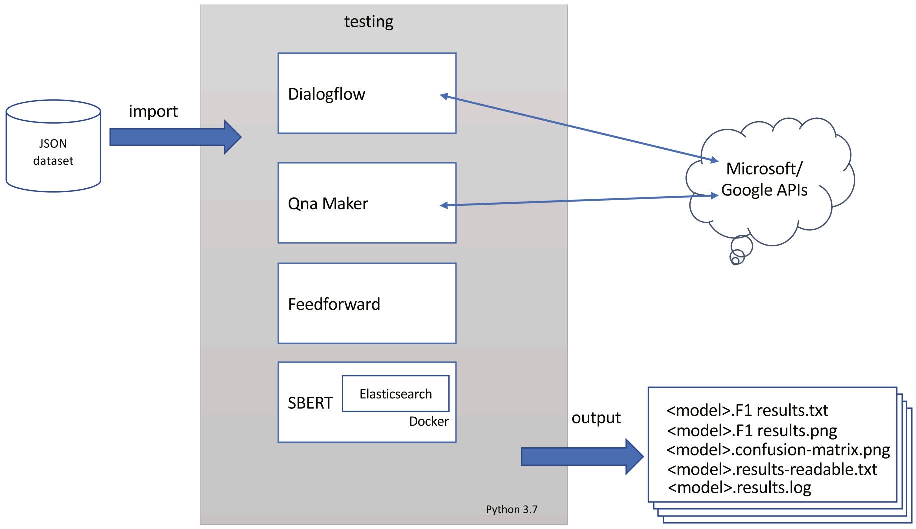
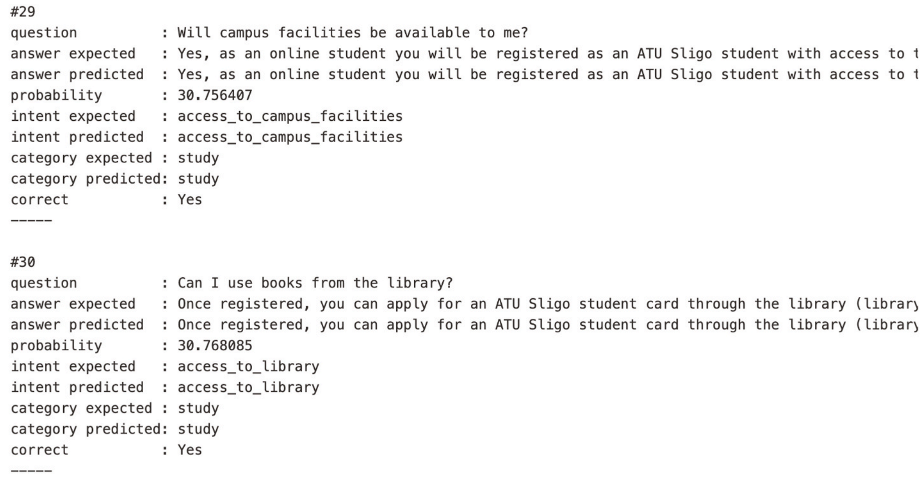


Fig. 5. Simple test harness using common processes for import, testing and output of results.



```
#29
question          : Will campus facilities be available to me?
answer expected   : Yes, as an online student you will be registered as an ATU Sligo student with access to t
answer predicted  : Yes, as an online student you will be registered as an ATU Sligo student with access to t
probability       : 30.756407
intent expected   : access_to_campus_facilities
intent predicted  : access_to_campus_facilities
category expected : study
category predicted: study
correct           : Yes
-----

#30
question          : Can I use books from the library?
answer expected   : Once registered, you can apply for an ATU Sligo student card through the library (library
answer predicted  : Once registered, you can apply for an ATU Sligo student card through the library (library
probability       : 30.768085
intent expected   : access_to_library
intent predicted  : access_to_library
category expected : study
category predicted: study
correct           : Yes
-----
```

Fig. 6. Sample readable, logging data.

Table 4
Comparison of model test results.

Model	Precision	Recall	F1-score
Feed Forward	0.64	0.60	0.56
Dialogflow	0.96	0.96	0.96
QnA Maker	0.96	0.96	0.95
SBERT	0.99	0.99	**0.99**

undergo changes and upgrades in the interim period since experimentation was completed. Secondly, datasets and the intents contained within may be from completely different domains which may not be directly comparable. Finally, the dataset in this study while complete for the specific use case, only contained 294 training phrases and 86 testing phrases. This appears to be very small in comparison with other studies where open domains have been utilised.

Notwithstanding the small dataset, the results of this study are very encouraging and suggest that further study is required using bigger datasets and multiple cross-validation techniques. This study would include investigations into newer, improved, pre-trained models rather than the “bert-base-nli-mean-tokens” model which has been recently deprecated. There are currently three models on the SBERT website that appear to have been specifically tuned for semantic search including the “multi-qa-mpnet-base-dot-v1”, “multi-qa-distilbert-cos-v1” and “multi-qa-MiniLM-L6-cos-v1” models. Consideration should also be given to using AI eco-systems like that provided by Hugging Face[10] which along with access to the models just mentioned as well as models from other users, also offers direct integration with testing tools from Weights and Biases.[11]

5. Conclusions

This study sought to establish where the highest NLU accuracy would be achieved by carrying out a comparative analysis between two popular chatbot frameworks with a Feedforward model, and an SBERT model in answering FAQs. A methodology was outlined whereby an FAQ dataset, associated with queries from prospective students applying to study online, was prepared and formatted for evaluation. Finally, an implementation was described which utilised a simple test harness which optimised and streamlined the creation of results.

It is intriguing to note the performance of the SBERT model compared to the Dialogflow and QnA Maker platforms in this work.

[10] https://huggingface.co.
[11] https://wandb.ai.

Existing platforms offer excellent environments for the development and delivery of a chatbot solution, while SOTA models can offer the potential of improved NLU accuracy. The user may wish to consider what the trade-off might be for their specific use case.

As the performance of the SBERT model is encouraging, further investigation would be required in a number of distinct areas. The fact that AI ecosystems are providing fast-evolving environments for the evaluation and comparison of language models has already been mentioned. Much of the comparative analysis already carried out uses datasets that are generic and limited in size. Since this study was completed, access has been gained to a Technological University online chat corpus from a Technological University. Comprising many thousands of interactions, this dataset may have dual potential in allowing the refinement of the existing FAQs as well as potentially enabling further testing of these FAQs with real questions.

Credit author statement

Kevin Peyton: Investigation, Writing – original draft, Writing – review & editing, Project administration Saritha Unnikrishnan: Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] M.J. van der Goot, T. Pilgrim, Exploring age differences in motivations for and acceptance of chatbot communication in a customer service context, in: A. Følstad, T. Araujo, S. Papadopoulos, E.L.-C. Law, O.-C. Granmo, E. Luger, P.B. Brandtzaeg (Eds.), Chatbot Research and Design, Cham, vol. 2020, Springer International Publishing, 2020, pp. 173–186.

[2] S. Yang, C. Evans, Opportunities and challenges in using AI chatbots in higher education, in: Proceedings of the 2019 3rd International Conference on Education and E-Learning, Association for Computing Machinery, Barcelona, Spain, 2019, pp. 79–83, https://doi.org/10.1145/3371647.3371659, available:.

[3] S. Cunningham-Nelson, W. Boles, L. Trouton, E. Margerison, A Review of Chatbots in Education: Practical Steps Forward [Chapter in Book, Report or Conference Volume], Engineers Australia, 2019.

[4] D. Braun, A.H. Mendez, F. Matthes, M. Langen, Evaluating natural language understanding services for conversational question answering systems, in: Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, 2017, pp. 174–185.

[5] C. Wisniewski, C. Delpuech, D. Leroy, F. Pivan, J. Dureau, Benchmarking Natural Language Understanding Systems, 2017.

[6] X. Liu, A. Eshghi, P. Swietojanski, V. Rieser, Benchmarking Natural Language Understanding Services for Building Conversational Agents, 2019 arXiv preprint arXiv:1903.05566.

[7] N. Malamas, K. Papangelou, A.L. Symeonidis, Upon Improving the Performance of Localized Healthcare Virtual Assistants', in Healthcare, MDPI, 2022, p. 99.

[8] P. Loeber, available: https://www.python-engineer.com/courses/pytorch beginner/13-feedforward-neural-network/, 2020. (Accessed 17 May 2021). accessed.

[9] N. Reimers, I. Gurevych, Sentence-bert: Sentence Embeddings Using Siamese Bert-Networks, 2019 arXiv preprint arXiv:1908.10084.

[10] N. Reimers, SentenceTransformers Documentation, 2016 available: https://sbert.net. (Accessed 18 January 2021).

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018 arXiv preprint arXiv:1810.04805.

[12] HEA, HEA - Springboard+, 2022 available: https://springboardcourses.ie/. (Accessed 18 August 2022).

[13] Technical Committee 39, ECMA 404 - the JSON Data Interchange Syntax, 2017 available: https://www.ecma-international.org/publications-and-standards/standards/ecma-404/. (Accessed 18 August 2022).