



CAN UNCLASSIFIED



DRDC | RDDC
technologyscience**technologie**

Question Answering Artificial Intelligence Chatbot on Military Dress Policy

A Natural Language Processing Based Application

Charith Gunasekara
DRDC – Centre for Operational Research and Analysis

Noah Chalifour
Matt Triff
Assistant Deputy Minister (Data, Innovation, Analytics) (ADM DIA), National Defence

Terms of Release: This document is approved for public release.

Defence Research and Development Canada

Scientific Report

DRDC-RDDC-2021-R175

November 2021

CAN UNCLASSIFIED

CAN UNCLASSIFIED

IMPORTANT INFORMATIVE STATEMENTS

This document was reviewed for Controlled Goods by Defence Research and Development Canada (DRDC) using the Schedule to the *Defence Production Act*.

Disclaimer: This publication was prepared by Defence Research and Development Canada an agency of the Department of National Defence. The information contained in this publication has been derived and determined through best practice and adherence to the highest standards of responsible conduct of scientific research. This information is intended for the use of the Department of National Defence, the Canadian Armed Forces (“Canada”) and Public Safety partners and, as permitted, may be shared with academia, industry, Canada’s allies, and the public (“Third Parties”). Any use by, or any reliance on or decisions made based on this publication by Third Parties, are done at their own risk and responsibility. Canada does not assume any liability for any damages or losses which may arise from any use of, or reliance on, the publication.

Endorsement statement: This publication has been peer-reviewed and published by the Editorial Office of Defence Research and Development Canada, an agency of the Department of National Defence of Canada. Inquiries can be sent to: Publications.DRDC-RDDC@drcd-rddc.gc.ca.

© Her Majesty the Queen in Right of Canada (Department of National Defence), 2021

© Sa Majesté la Reine en droit du Canada (Ministère de la Défense nationale), 2021

CAN UNCLASSIFIED

Abstract

In the Canadian Armed Forces (CAF) there are many policies and standards that must be followed. These policy documents can be quite long and sometimes difficult to navigate. For example, the Department of National Defence (DND)/CAF corporate policies in [1] are available in nested weblinks with seemingly overlapping titles with various chapters; each comprises hundreds of paragraphs with no direct search option strictly for the policy document. In a situation where someone needs information that may be in a policy, it can take quite a bit of time to get an answer when scrolling through a very comprehensive policy or standard document. This Scientific Report explores how artificial intelligence (AI) using natural language processing (NLP) can be used to develop a question answering chatbot that can retrieve passages from a policy to answer a user's question.

We first develop a data collection method to automatically read the policy and standards documents and then implement an NLP based data retrieval system to find the most suitable answer to users' questions. We also introduce a mechanism for the chatbot to self-learn using the feedback sent by the users, allowing it to adaptively change its algorithm to improve the accuracy. The chatbot was tested against sample questions sent by CAF members and early testing shows a high accuracy in the responses.

Significance to Defence and Security

Using an AI-based question answering system to provide relevant answers to policy questions would allow CAF members to have a simple chat interface where they can get an answer to their questions instead of taking the time to look through the policy or standard document themselves. This would result in less time wasted looking through long policy documents as well as the potential of false information through miscommunication when asking other members of the CAF. The chatbot will also reduce the traffic of questions sent to policy committees, minimizing disruptions to the committee members' other critical duties.

Résumé

Dans les Forces armées canadiennes (FAC), de nombreuses politiques et normes doivent être respectées. Les documents énonçant ces politiques et normes peuvent se révéler très volumineux, et il est parfois difficile de s'y retrouver. Par exemple, les politiques ministérielles du ministère de la Défense nationale (MDN)/CAF dans [1] sont accessibles sous forme de liens Internet imbriqués avec des titres qui semblent se chevaucher. Ils comprennent plusieurs chapitres et chaque chapitre contient des centaines de paragraphes. Il n'existe aucune fonction permettant d'effectuer une recherche directe ciblée dans ledit document. Par conséquent, lorsqu'une personne doit rechercher un renseignement, il lui faut parfois beaucoup de temps pour parcourir ce document exhaustif avant d'obtenir une réponse. Ce rapport scientifique explore comment l'intelligence artificielle (IA) utilisant le traitement du langage naturel peut être utilisée pour développer un agent conversationnel qui peut trouver les passages pertinents dans une politique pour répondre à la question d'un utilisateur.

Nous développons tout d'abord une méthode de collecte de données pour lire automatiquement les documents relatifs aux politiques et aux normes, puis nous mettons en œuvre un système d'extraction de données basé sur le traitement du langage naturel pour trouver la réponse la plus appropriée aux questions des utilisateurs. Nous introduisons également un mécanisme permettant à l'agent conversationnel d'apprendre par lui-même en utilisant le retour d'information envoyé par les utilisateurs, ce qui lui permet de modifier son algorithme de manière adaptative pour améliorer la précision. L'agent conversationnel a été testé sur des exemples de questions envoyées par les membres des FAC et les premiers tests démontrent une grande précision dans les réponses.

Importance pour la défense et la sécurité

L'utilisation d'un système de réponse aux questions basé sur l'IA pour fournir des réponses pertinentes aux questions sur les documents relatifs aux politiques permettrait aux membres des FAC de disposer d'une interface de clavardage simple où ils pourraient obtenir une réponse rapide à leurs questions au lieu de prendre le temps de parcourir eux-mêmes les documents relatifs aux politiques ou aux normes. Il en résulterait une économie de temps importante pour les militaires qui n'auraient plus à examiner ces documents volumineux. De plus, le risque d'obtenir des renseignements erronés en raison d'une mauvaise communication avec les autres membres des FAC serait grandement réduit. En outre, le robot conversationnel permettra également de réduire le nombre de questions envoyées aux comités d'orientation, ce qui minimisera les perturbations des autres tâches essentielles réalisées par les membres de ces comités.

Table of Contents

Abstract	i
Significance to Defence and Security.	i
Résumé	ii
Importance pour la défense et la sécurité	ii
Table of Contents	iii
List of Figures	iv
List of Tables	v
Acknowledgements	vi
1 Introduction	1
1.1 Background	1
1.2 Previous Work	1
1.3 Outline	2
2 Data Processing	3
2.1 Policies and Standards Data	3
2.2 Data Collection	3
2.3 Data Structure	5
3 Methodology	6
3.1 Answering Algorithms	6
3.1.1 Early Experimentation	6
3.1.2 Final Algorithm	9
3.2 User Interface (UI) Development	11
3.3 Deployment	12
3.4 System Evaluation	14
3.5 Model Fine-tuning	14
4 Performance Evaluation	15
5 Future Work	27
6 Conclusion	28
References	29
List of Symbols/Abbreviations/Acronyms/Initialisms.	32

List of Figures

Figure 1:	Block diagram of the data collection algorithm.	4
Figure 2:	Data structure of knowledge base and dialogue management.	5
Figure 3:	BERT encoder [14] (t_i represent the input tokens, w_i represent the encoded values from transformer and t'_i represent the values normalized by classification layer using softmax function).	7
Figure 4:	Transformer architecture.	8
Figure 5:	Multi-head attention in transformer models.	9
Figure 6:	Bi-encoder architecture overview.	10
Figure 7:	Cross-encoder architecture overview.	11
Figure 8:	Web UI built with React.	11
Figure 9:	A comparison of both simple and scalable deployment models.	13

List of Tables

Table 1: Test questions and answers.. 16

Acknowledgements

The authors would like to acknowledge the members of the Canadian Armed Forces including Chief Warrant Officers who provided feedback and test questions and answers to improve the performance of the chatbot; Maj David Dunwoody for serving as the project manager; Mr. Steven Horn, Director of Data Science at ADM(DIA), Dr. David Waller, Section Head, Materiel Infrastructure, Economics and Analytics, Defence Research and Development Canada (DRDC) – Centre for Operational Research and Analysis (CORA) and Mr. Yvan Gauthier, Former Director of Data Science at ADM(DIA) for providing funding and resources for the project.

1 Introduction

1.1 Background

Policy documents contain plenty of valuable information that Department of National Defence (DND) and Canadian Armed Forces (CAF) personnel must follow. These documents can be long and difficult to navigate, making it hard to find information within them. The CAF have various policy and standards documents that must be adhered to by all its members. As of right now, most questions regarding the policy and standards documents are answered by various committees. This process usually involves sending an email to the committee in question and awaiting a response which could take some time as answering questions is a secondary task to the members of committees. Therefore, it would be beneficial to have an automated tool to answer questions from policy and standard documents. Such a tool would allow for fewer inquiries to the committees and eliminate waiting time before an answer is received.

Question answering systems have proven useful in information retrieval tasks with the implementation of modern Natural Language Processing applications. This study aims at applying state-of-the-art question answering algorithms in a military policy and standard document setting. Our goal is to develop a system where a user can ask a question using natural language, and our system will retrieve the passage it believes contains an answer to the user's question.

1.2 Previous Work

Machine learning is the study of computer algorithms that improve over time and iteration of data. The purpose of these machine learning algorithms is to give them some input data, like an image of a written number, and have the algorithm provide an output, in this case, the number written in the image. These algorithms use statistical analysis to derive a mathematical formula that will produce output data when applied to the input data. Basic machine learning models have proven very useful, but they have their limitations and cannot provide value when the problem they are trying to model gets too complicated. Deep learning aims at building on top of machine learning. In essence, it uses multiple machine learning algorithms stacked on top of each other in "layers" and form what is called a neural network. By stacking these algorithms, the mathematical formulas that can be derived become a lot more complex, and it enables the computer to find solutions for much more complicated problems [2].

Natural language processing (NLP) is the study of algorithms that can help computers understand and interpret human language. The objective of NLP is to be able to use machine learning with human language. Since machine learning works by deriving a complex mathematical formula for the problem, it only works with numbers and cannot be used with textual data. There are a variety of NLP techniques that convert a body of text into a numerical array or matrix. One of the most basic NLP techniques for text to number conversion is creating a word vocabulary. By doing this, the algorithm can associate each unique word in the vocabulary with a unique number. Using this technique allows the algorithm to take a body of the text of any size and convert it into a list of numbers which it can use as a matrix. This type of process of converting text to numbers is known as encoding, and it is one of the base principles of NLP [3][4].

NLP and machine learning have been combined to create many practical applications. Google's search engine is one of the most known applications of NLP. Google uses NLP to match a user's query with a set of websites that are relevant to the query [5]. The search engine splits the user query into separate words at

an elementary level and then searches for those words in websites and then ranks them by the most relevant based on those keywords. Another more advanced usage of NLP embedded into Google's search engine is question answering. Question answering is a much more complex NLP problem where the algorithm doesn't just try to find webpages that contain relevant information but tries to extract the exact answer from a body of text (such as a webpage). This feature is demonstrated when you search a question in Google; a lot of the time, it will show a card as the first result with a body of text, and some of the text is bolded, which is the extracted answer. The techniques used for question answering are relatively more complex than a keyword search, and they use deep learning models such as Bidirectional Encoder Representations from Transformers (BERT) [6] to encode the text into a more meaningful matrix.

While a search engine like Google is an extensive implementation of question answering, there have also been some more specific applications. Policy documents such as privacy policies for companies are very complex documents that are long and messy. Experiments such as [7] conducted by Carnegie Mellon University, Penn State University and Fordham Law school, have shown that question answering techniques can be used for users to get answers about a policy document. In their experiment, they used various models including BERT to benchmark the PrivacyQA dataset. Given the complexity of policy documents such as privacy policies, the ability to ask questions about a policy document would be incredibly beneficial. The results in [7] achieved a precision of 44% vs. the 69% of a human question answering baseline. These results are promising but leave a lot of room for improvement.

In this work, we apply the retrieval and reranking information retrieval (IR) technique from [8] to military policy and standards documents. A set of questions was curated in order to measure the performance and viability of the proposed question answering system.

1.3 Outline

The remainder of this Scientific Report is organized as follows. Section 2 describes the data and how it was curated and processed for the question answering system. Section 3 will go into detail about the model structure and the components that were used in our study. This section will also contain training details and the procedures used to fine-tune pre-trained models for military policies. Section 4 shows and analyzes the results of the system on a collection of validation questions. Section 5 discusses future work for further improvement of the chatbot. Finally, the Report is concluded in Section 6 with a summary of all the key findings from our study.

2 Data Processing

2.1 Policies and Standards Data

The data used in this study was the National Defence Policies and Standards, which can be found at [1]. All the data is stored on web pages which is not ideal for usage. The policies are formatted using bullet points, where bullets can have sub-bullets that contain additional information about their parent bullet. There are also a variety of tables in the policies and standard web pages. Unfortunately, however, many of the policies are not in the same format, making it more challenging to parse.

2.2 Data Collection

Given that all the data is stored on a webpage, we needed to build a way of automatically parsing it and storing it in some database. We decided it would be best to create a variety of scripts that can be run in order to extract all the data when executed. By creating separate scripts for data extraction, we can run the scripts at any time, which means when the chatbot is deployed to production, we can add the parsing scripts to a job that is scheduled, for example on a daily basis. Most of the policies are formatted differently, which means we needed to create separate scripts for each unique policy format. In order to speed up the development and experimentation of the core chatbot, we decided to focus on the Canadian Forces Dress Instructions [9] at first and then expand from there.

To parse all the data, we built several Python scripts that use the BeautifulSoup package [10] to scrape the data from the web pages. The main format for the policies was bullet points, so to scrape that, we built a recursive algorithm that would start with the headers for each section and then for each bullet point it would go all the way down to the “leaf” bullet point, which is the bullet that has no sub-points. The scraper would then save the text of the bullet point along with the HyperText Markup Language (HTML) to use the text for the question answering system and the HTML for displaying to the user. Figure 1 provides a block diagram of the data collection algorithm.

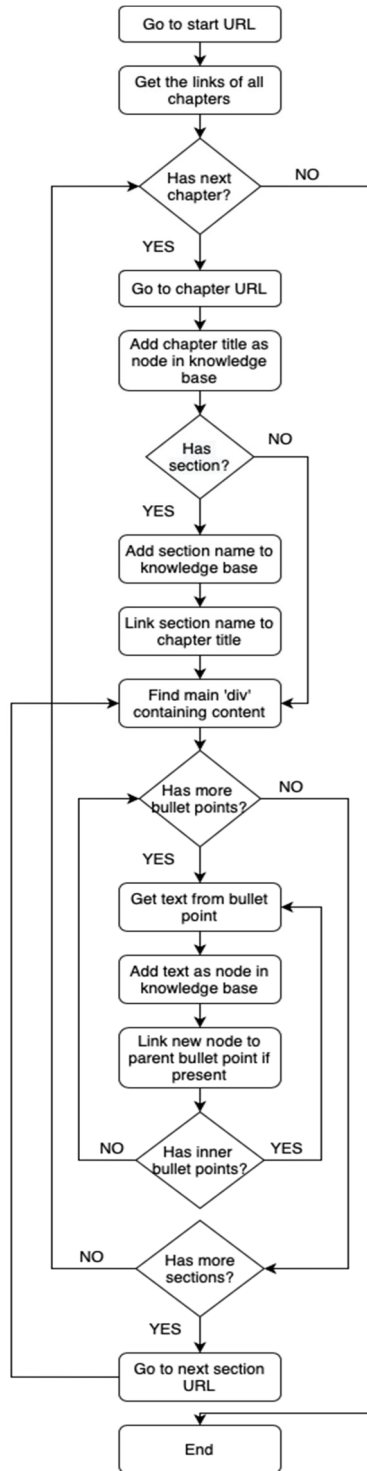


Figure 1: Block diagram of the data collection algorithm.

2.3 Data Structure

In order to store all the data in a database for it to be queried later, we needed to design a data structure that will allow us to reference the text in the policy documents and the format of the document. Since most of the policy documents are stored in bullet points with some points having sub-points, we decided that a tree would be a good structure since it closely resembles the bullet point structure. We call this tree data structure the knowledge base. The knowledge base is composed of nodes and edges. Each node represents a bullet point from the policy document and stores all the relevant information about the text. The edges represent connections between the nodes and subpoints to form a tree structure. Each edge has a parent and child to know each node's level in the tree. In our database, the tree data structure has nodes in the following order connected by edges, policy document title, chapter name, subtitle, table title (if any), policy passage, and sub-points in the policy passage (if any).

Along with all the policy information, we also want to store all the chatbot dialogues with its users. We store these dialogues so that we can fine-tune the chatbot and improve the passage retrieval models. To store all these dialogues, we use sessions and utterances. The session represents the entire dialogue between a user and the chatbot, and the utterance represents a single message in the dialogue. See Figure 2 for the complete design of the data structure. The data structure is stored in SQLite data format [12].

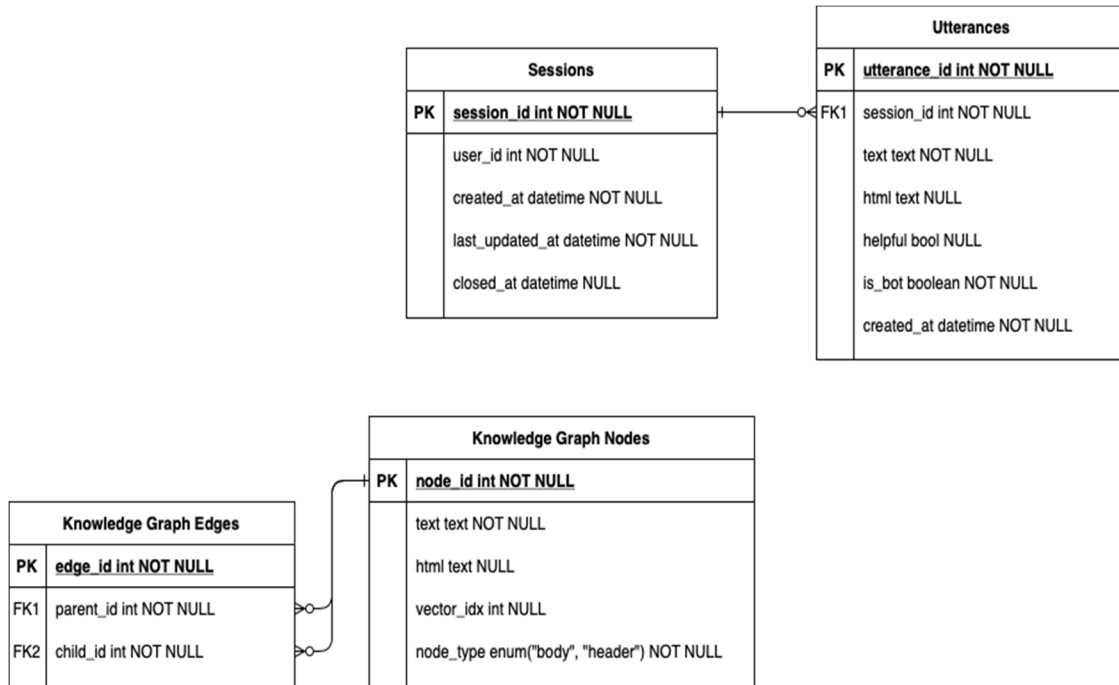


Figure 2: Data structure of knowledge base and dialogue management.

3 Methodology

For a question answering system to provide value to its users, it needs to give correct answers to a question; if not, it could confuse its users. To remove the ability to provide generated answers outside the document corpus, we developed a retrieval-based chatbot instead of a generative-based chatbot. In retrieval-based question answering, all of the answers can be found directly in the knowledge corpus, whereas in generative-based (for e.g., GPT-3 [11]), it generates unique answers on its own. This will ensure that the chatbot does not give information that is not directly mentioned in the policy documents to the user which could cause more confusion than it provides value.

3.1 Answering Algorithms

3.1.1 Early Experimentation

Throughout the experimentation and validation of the question answering chatbot, many answering algorithms were evaluated to find the one with the highest accuracy to give the best possible experience to its users. We define an answering algorithm as an algorithm composed of NLP techniques and machine learning to extract a passage from a policy document.

All our experiments have been built around the idea of using BERT to encode text into a vector and using this technique to retrieve the most relevant passage to our question. BERT is a language representation model which uses a Transformer model [13] instead of the more common recurrent neural network in language modelling. Transformer models are based solely on attention mechanisms instead of using recurrent networks [15]. The self-attention mechanism allows transformer-based algorithms to build connections between the meaning of words in a sentence. Whereas traditional deep learning based models such as Recurrent Neural Networks (RNN) rely on constructing the meaning-based order of the words that appear in a sentence [15]. Since a sentence can be written in different ways by changing the word order, RNN often fails to identify the sentences with the same meaning if words are structured differently. Therefore, self-attention allows the Transformer model to learn much quicker and decreases the complexity of the language model. The key feature of BERT is that the model is trained bidirectionally. When the model encodes a word, it does not just look at the previous words like in traditional recurrent language models. When encoding a sentence using BERT, the model encodes each word into an embedding context-aware from both sides, meaning it considers all the words before and after it in the text.

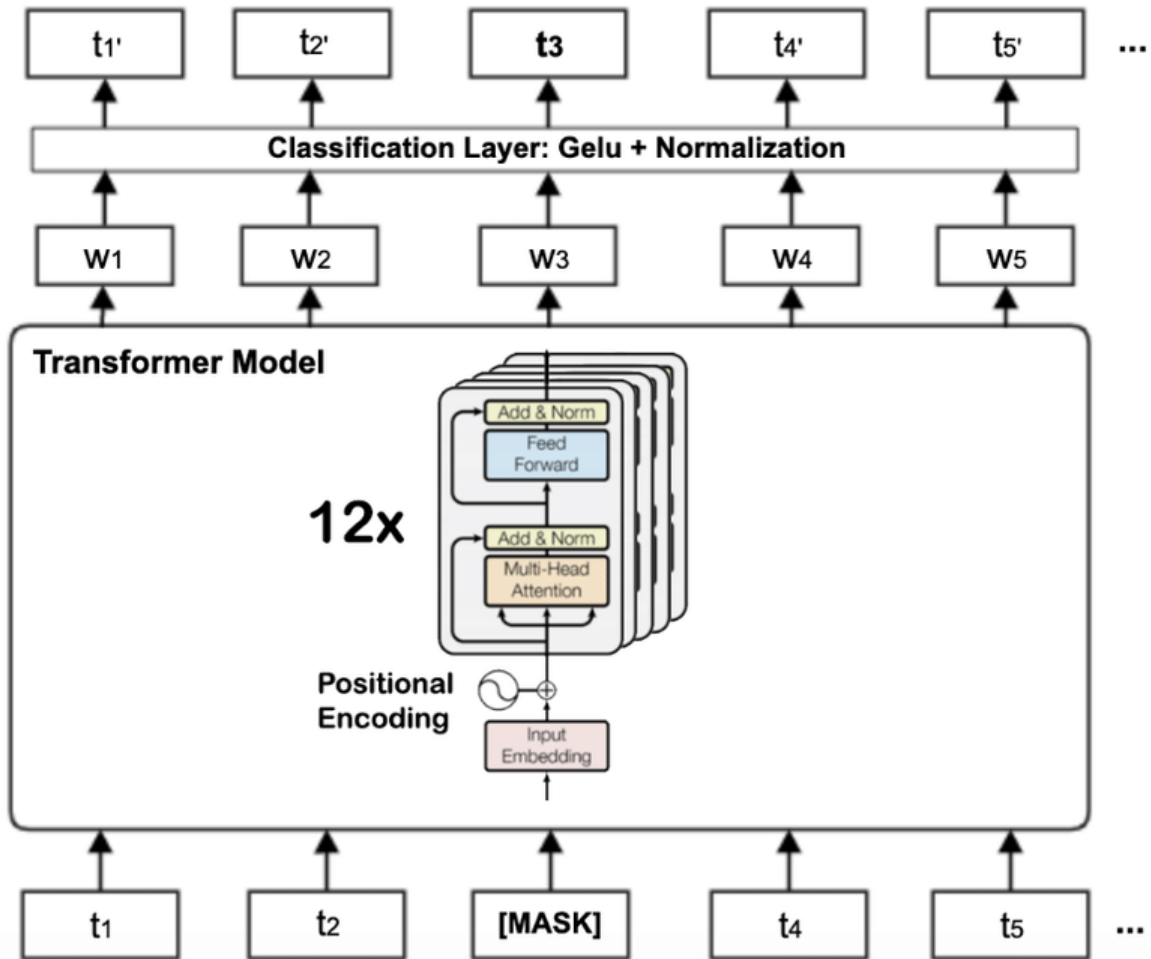


Figure 3: BERT encoder [14] (t_i represent the input tokens, w_i represent the encoded values from transformer and t'_i represent the values normalized by classification layer using softmax function).

As you can see in Figure 3, BERT relies on the encoder from the Transformer model. The Transformer, shown in Figure 4, initially proposed by Google was meant as an encoder-decoder architecture for sequence-to-sequence modeling; BERT is intended to be used only as a text encoder for tasks such as classification or, in our case, question answering. The Transformer's encoder is built using multiple "encoder blocks," and in the case of BERT, it uses six blocks. Each of these encoder blocks are composed of a multi-head attention component and a basic feed-forward network.

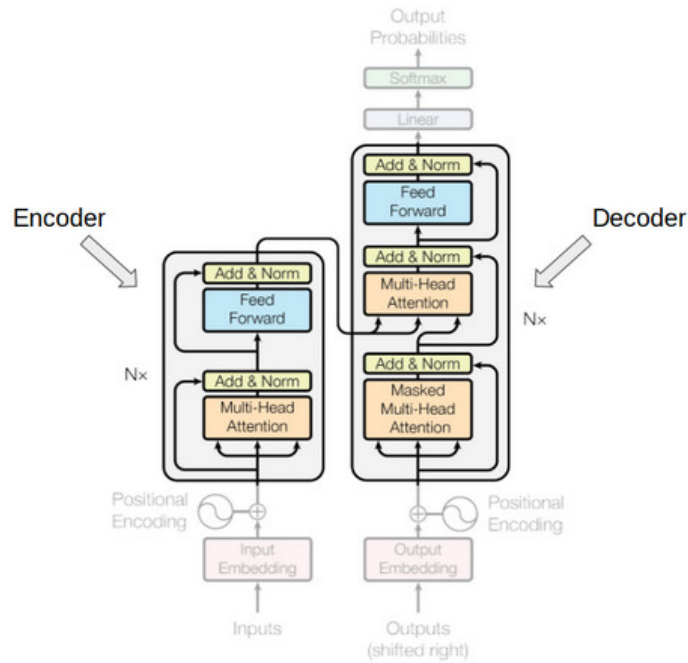


Figure 4: Transformer architecture. Reproduced with permission from “A. Vaswani.”
The original can be found in [15].

The multi-head attention mechanism shown in Figure 5, uses multiple “heads” of dot-product attention and concatenates the resulting vectors together. The transformer model first establishes three vectors: query, key, and value vectors. The query vector represents the relative importance of the words in a question for a given output passage defined by the key vector; the value vector calculates the weight of the relevance of each word in the output passage to the input query (the question). All these three vectors are pre-trained in the BERT model [15]. The dot product is then taken between the query and the key vectors. The resulting values are then normalized using the softmax function¹ and multiplied with the values vector. This allows for the model to retain more information about dependencies between tokens in the text. In the initial Transformer model described by Google, they decided to use eight heads for the attention mechanism, which showed a significant performance improvement.

¹ The softmax function is a function that converts a vector of K values into vector of K values that sum to 1 which can be interpreted as probabilities [17].

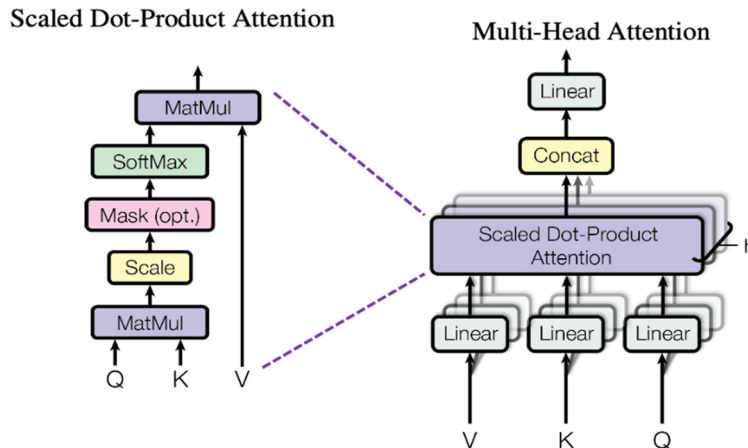


Figure 5: Multi-head attention in transformer models. Reproduced with permission from “A. Vaswani.” The original can be found in [15].

In our initial experiments we tried simply using a small BERT model [19] to encode all the text from the policy documents and the questions posed to the chatbot, and then use a cosine similarity function to find the most relevant passage to the question. A smaller BERT model was chosen for fast inference speeds to get answers in real-time. In doing so, we realized that this model does not work well in the context of question-answering as there are many more factors other than relevance that determine if a passage contains the answer to a question. It also showed us that a single model might not be optimal for our question-answering task since there are thousands of possible passages.

To solve the problem of too many possible passages for a single model to retrieve from, in our next experiment, we tested using a simple, fast bag-of-words ranking function called BM25 [20]. We tested using this function to do a quick filter of all the thousands of passages and retrieve the top 100 to then feed into BERT to retrieve the best possible passage to answer the question. The BM25 function has been shown to be very useful for information retrieval therefore we wanted to explore the feasibility of using it for question answering. This experiment showed us that a retrieve and rank algorithm would perform better for our policy passage retrieval; however, the BM25 and small BERT models we were using were not designed to answer questions.

3.1.2 Final Algorithm

The following few paragraphs detail the retrieve and rank algorithm that we implemented and the models that power it. Since there are thousands of possible passages that can contain the answer to a question, we needed to first implement an algorithm that can encode passages into a vector and work in a question answering setting. This vectorization is an essential part of any NLP model to convert words into a numerical representation for the computer to process. We ended up deciding on the msmarco-distilbert-base-v3 model from [21]. This encoder model has been pre-trained on the MSMARCO Passage Ranking dataset [22]. We use this model as our encoder, therefore when we parse the policy documents using our scrapers, we encode each passages using this model and store it into a large matrix for later retrieval. In a bi-encoder, shown in Figure 6, the encoder takes the question and the candidate passage and calculates question embedding and passage embedding separately. The encoder first splits the text into a set of tokens then uses a transformer model to encode all the tokens with a special token prepended into a set of vectors. A reduction function is

then used to convert the set of vectors into a single vector. In our case, we use the vector of the special token as a representation of the entire text. The special tokens are unique numbers added to surround an input text passage for the model to easily find the beginning and the end of the given text [23].

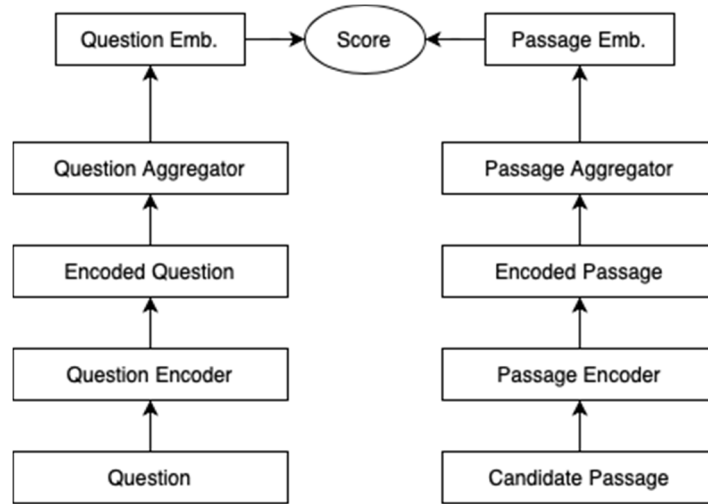


Figure 6: Bi-encoder architecture overview.

Once we have encoded all our passages, we needed to find some way of filtering these encodings based on our question. To do this, we used a Python package called `sentence_transformers` [24], which has a function called `semantic_search`. This function performs a cosine similarity between our question vector and all the passages.

With the top k most relevant passages extracted, we needed to develop a way of re-ranking the passages based on which is most likely to answer the question. To do this, we used a cross-encoder [3][4]. A cross-encoder is a model that embeds the question and passage texts together and computes a score. We use the cross-encoder to rank all the passages extracted by the bi-encoder. We use it here because it has been shown to efficiently extract the correct passage as opposed to the bi-encoder; however, we cannot use the cross-encoder on the entire corpus as it is very inefficient on large amounts of text. The model was trained for question answering in our context, so the score signifies the likelihood that the passages answer the question. When using a cross-encoder model, the encoder embeds the question and candidate passage together, allowing it to consider the interaction between the input and candidate passage, meaning every word in the candidate passage can attend to the words in the input text and perform self-attention between the question and passage. By doing this, the cross-encoder can achieve richer attention and better distance computation between the input text and candidate passage than the bi-encoder. To do this, before we encode the text, we concatenate the question and passage together with a special token between them as in [23]. In our implementation, we used the cross-encoder/ms-marco-TinyBERT-L-4 model from [21], which was made possible with the `sentence_transformers` package [24].

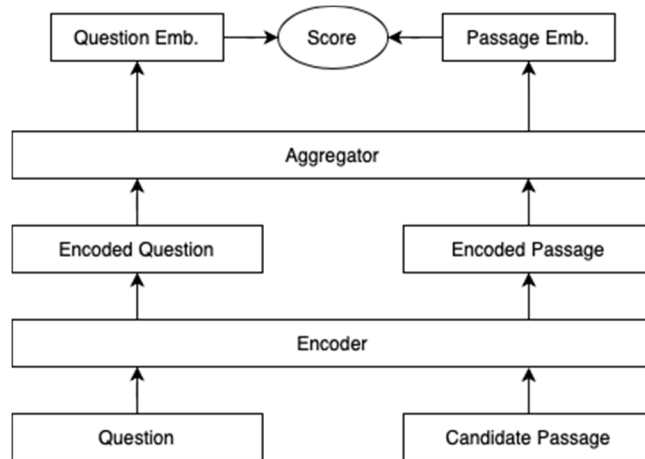


Figure 7: Cross-encoder architecture overview.

Once we have ranked the passages, to select the answer to send back to the user, we take the passage with the highest score from the cross-encoder containing the answer and return that to the user.

3.2 User Interface (UI) Development

To get some feedback early in the experimentation process, we decided it was best to build a web application where users could go and test out the chatbot. We used React [26] as a frontend framework for the user interface (UI) development. For the web application backend, we used Flask [27] as it is a very lightweight basic application programming interface (API) framework that suited the needs of the application.

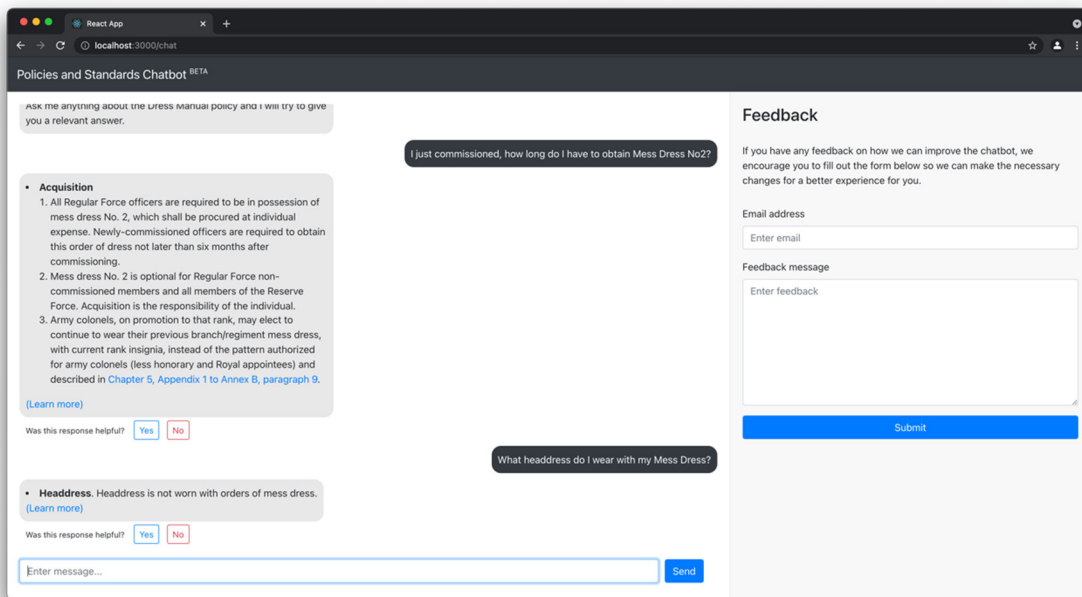


Figure 8: Web UI built with React.

Microsoft Teams is a common form of communication throughout the organization. Therefore, we developed a gateway to communicate with the chatbot through Teams using the Microsoft Bot Framework [28].

3.3 Deployment

The chatbot has so far been run in a local development environment for demonstration and testing purposes. However, for the chatbot to be used throughout DND/CAF, it must be deployed and hosted to be accessible.

To facilitate deploying the application, we have split it into three components. A front-end web application, a back-end API server, and a database. To facilitate easy and scalable deployment of the tool, both the front-end and back-end have been separated into their own Docker [29] containers. Containers provide the benefit of isolating application components from underlying system dependencies and requirements. This allows applications to be repeatedly deployed without significant setup on the servers from which they are run.

The back-end API server is the most computationally demanding component of the chatbot. This component processes questions from users and determines the most appropriate answer, as described in Section 3.1. The front-end web application consists of the interface described in Section 3.2. The front-end is relatively lightweight and does not require significant computing power to run. Currently, a SQLite database [30] is used within the back-end container to store the policy corpus, as well as previous questions and feedback from users.

Since the chatbot application uses only publicly available policy information, it is a good candidate for deployment in a cloud environment. The most straightforward deployment of the system could use an infrastructure-as-a-service (IaaS) offering from a cloud provider. A cloud virtual machine (VM) with Docker installed could run both the front-end and back-end containers (the back-end container also includes the SQLite database). Access to this VM could then be granted to users who would use the web interface to ask questions. Figure 9a shows a system diagram for this deployment model.

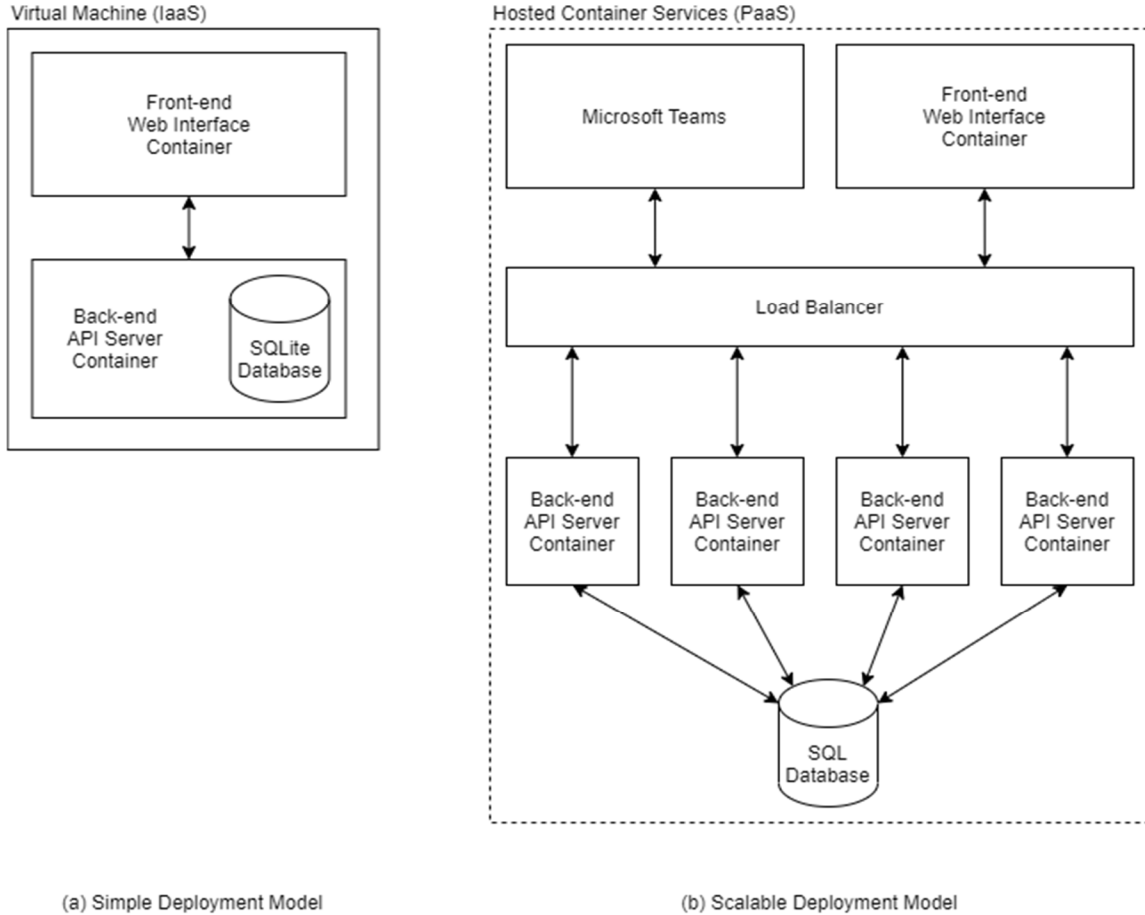


Figure 9: A comparison of both simple and scalable deployment models.

However, the simple deployment approach is not appropriate for a system that is deployed to the entire DND/CAF organization. It would not be able to process a large number of simultaneous requests. Instead, the following changes should be implemented. The back-end API server is the most computationally intensive and thus the most likely bottleneck to the system. However, we can leverage Docker containers to deploy multiple copies of the back-end. To do so, the SQLite database would be replaced with a separate, hosted SQL database. All instances of the back-end container would connect to this centralized database. Next, a load balancer component would be introduced. When a new session is opened with the chatbot, the load balancer will assign that session to a specific instance of the back-end and ensure that no instance has more requests than it can handle. Finally, the front-end web interface may not be the best (or only) method users will want to connect with the system. Microsoft Teams can be used as the front-end in addition to the web-based interface. All of the above components could then be deployed using a platform-as-a-service (PaaS) offering from a cloud provider. These offerings do not require administrators to interact with specific VMs, resulting in fewer resources needed to maintain the system. Figure 9b shows a system diagram for this deployment model.

3.4 System Evaluation

In order to evaluate our chatbot we decided to gather a test bank of questions and the passage that contains the best answer to the best of our knowledge.

To create a test bank of questions, we demoed our application early in the testing phase to get our app in front of users as early as possible and get some real potential questions that they are asking. During these demos [31][32][33], we were able to gather a number of questions that we used to get an early sense of the chatbot's performance.

Once we obtained enough question-answer pairs, we created a script that would evaluate the performance of the chatbot. Our model would go through each question in the question bank and compare its output with the expected output from the test bank. We then calculated the number of questions the model answered correctly. To give our evaluation a metric, we divided the number of questions answered correctly by the total number of questions in the test bank to get a percentage of questions answered correctly. We decided that this metric is ideal to evaluate the performance of the chatbot and its success rate once deployed.

We also built an answer feedback mechanism into the UI under each response from the chatbot where the users can tell the chatbot if the answer provided was helpful. We also use this feature to evaluate the chatbot by measuring the percentage of questions that the chatbot can answer correctly.

3.5 Model Fine-tuning

In order to improve the performance of the chatbot, we decided to experiment with fine-tuning the BERT encoder used in the model. Model fine-tuning refers to the idea of training a model on a dataset that is more specific to the task we are trying to accomplish. In our case, the two encoder models that we are using (bi-encoder and cross-encoder) were trained on a very general question-answering dataset, however, the policy documents that our chatbot needs to understand and answer questions about have more policy and military-specific wording. Therefore, it is quite possible that the performance can be improved by fine-tuning on a more policy or military-oriented dataset in order for the model to learn a more specific vocabulary and get a better understanding of the questions asked and the policies themselves.

We take a two-step process at fine-tuning our chatbot. First off, we want to fine-tune the bi-encoder model by training it on all our military policies and standards documents using the method described in [34]. We use standard BERT training to fine-tune the model to adapt to our military-specific vocabulary. This method uses the transformers package [34] to fine-tune BERT using input masks which means during training, we remove some words from the input text, and BERT tries to fill in the blanks.

Secondly, we want to fine-tune our cross-encoder model to improve our answer re-ranking in our answering algorithm. In order to do that we use the answer feedback from the users to our chatbot and apply it to Scenario 2 of [1] where we tune cross-encoder and bi-encoder iteratively comparing the question-answer pairs from the model with the dataset used for fine tuning. This method of fine-tuning takes the question-answer pair as input and a score as output where the score is in the range of 0 to 1 representing a percentage. In our case, we can use the message feedback when a user says whether a response was helpful or not as our output for our training data, 0 being not helpful and 1 being helpful. We are planning to do start this fine-tuning task as a future work once we have collected enough user feedback.

4 Performance Evaluation

This section describes the performance of our proposed chatbot model on our evaluation test bank of questions. In Table 1, we show the set of question-answer pairs from our test bank along with the answer retrieved by our chatbot model. Through our evaluation, we have found that our model successfully answers fundamental questions; however, it has difficulty when questions require more information. Especially questions 18 and 19 show how giving detailed information can help the chatbot retrieve the correct answer. Out of 26 questions, the chatbot was able to get the exact answer 21 times, and it failed to give the correct answer only for three questions. For one question, it gave the correct answer as the 4th best response (we gave 25% accuracy for that response), and for another question, the chatbot's answer was the paragraph above the expected answer; however, that paragraph was pointing towards the correct answer; therefore, we gave that response a score of 95%. Hence, we can calculate the accuracy of the chatbot using the current test bank follows.

$$Accuracy = \frac{21 \times 100 + 1 \times 95 + 1 \times 25 + 3 \times 0}{26} \times 100\% = 85.4\%$$

Table 1: Test questions and answers.

No.	Question	Expected answer	AI chatbot’s answer	Accuracy
1	At what height should the ranks be sewn on the tunic?	(1) Sewn centred on both sleeves with the inside point of the bottom chevron 9 cm above the bottom edge of the sleeve. Standard badges of rank shall be worn in their normal locations (see also Chapter 5, Section 2, paragraph 13), except for chief petty officer, petty officer first class and warrant ranks, nwhich shall be sewn centred 1.2 cm above the apex on the uppermost chevron.	(1) Sewn centred on both sleeves with the inside point of the bottom chevron 9 cm above the bottom edge of the sleeve. Standard badges of rank shall be worn in their normal locations (see also Chapter 5, Section 2, paragraph 13), except for chief petty officer, petty officer first class and warrant ranks, which shall be sewn centred 1.2 cm above the apex on the uppermost chevron. (Learn more) .	100%
2	Can the toque be worn with the Naval Wind Raingear?	Winter dress. parkas, Yukon caps, winter fur cap, toques and scarves may be worn only when winter dress is in effect. Operational Gore-Tex parkas may be worn with No. 3 orders of dress as follows: 1. Navy personnel—Navy Gore-Tex parka, NSN 8415-21-920-2414, and Naval Wind Raingear (jacket only); 2. Air Force personnel—intermediate jacket, NSN 8415-21-910-8219 and cold weather parka, NSN 8415-21-910-8195. <ul style="list-style-type: none"> • This is the sole instance where operational clothing may be worn with the No. 3 orders of dress. The optional Service Dress Parka and topcoat (gabardine) are authorized for winter outer wear when in service dress. The topcoat will remain the outer wear for parades and other ceremonial activities. 	Winter dress. parkas, Yukon caps, winter fur cap, toques and scarves may be worn only when winter dress is in effect. Operational Gore-Tex parkas may be worn with No. 3 orders of dress as follows: 1. Navy personnel—Navy Gore-Tex parka, NSN 8415-21-920-2414, and Naval Wind Raingear (jacket only); 2. Air Force personnel—intermediate jacket, NSN 8415-21-910-8219 and cold weather parka, NSN 8415-21-910-8195. <ul style="list-style-type: none"> • This is the sole instance where operational clothing may be worn with the No. 3 orders of dress. The optional Service Dress Parka and topcoat (gabardine) are authorized for winter outer wear when in service dress. The topcoat will remain the outer wear for parades and other ceremonial activities. (Learn more) .	100%

CAN UNCLASSIFIED

No.	Question	Expected answer	AI chatbot's answer	Accuracy
3	With which number of dress can I wear my beret?	The beret may be worn as an optional headdress with the following Navy order of dress: Nos. 3A, 3B (with black trousers/ pants/skirt) and 3C. The service dress CAP/HAT shall be the only authorized headdress to be worn with No. 3 and 3B (summer whites) orders of dress.	The beret may be worn as an optional headdress with the following Navy order of dress: Nos. 3A, 3B (with black trousers/ pants/skirt) and 3C. The service dress CAP/HAT shall be the only authorized headdress to be worn with No. 3 and 3B (summer whites) orders of dress. (Learn more) .	100%
4	When can I wear the Yukon hat?	headdress—shall be worn outdoors. Toques and Yukon caps may only be worn during winter dress period with the topcoat (gabardine), and DEU Parkas;	The tuque, navy, black or hat, winter black, Yukon is authorized for wear with the MPOPD in cold or inclement weather during winter dress periods. The MP Branch cap badge shall not be affixed to either form of winter headdress; and (Learn more) . <i>Note: In this case the chatbot pulled another passage with the same meaning as the expected answer</i>	100%
5	When does the dress code change from winter to summer?	Changes into summer and winter dress shall be ordered by commanders of commands and NDHQ group principals for all subordinates within Canada and its territorial waters. Overseas commanders may conform to local circumstances.	Changes into summer and winter dress shall be ordered by commanders of commands and NDHQ group principals for all subordinates within Canada and its territorial waters. Overseas commanders may conform to local circumstances. (Learn more) .	100%
6	Can I wear my uniform and medals to a Levee?	Formal and other significant occasions for which the wearing of complete ceremonial attire—No. 1 or 1B orders—is not deemed necessary or appropriate; i.e., no swords, ceremonial belts, bayonets, etc.: a. investitures; b. levees; c. ceremonial parades; d. ceremonial occasions, when attending as a spectator; e. on Remembrance Days in messes; f. formal military weddings; and	Civilian Morning Dress, Lounge Suits, Blazers, and Day Dresses 1. Where decorations are appropriate for the occasion, such as investitures, Remembrance Day and other memorial services, and some formal parades, one neck badge and/or full sized medals are normally worn. 2. As many as four stars of orders may be worn on a tailcoat, but only one on a director's (short black) coat or an afternoon dress. Stars of	0%

CAN UNCLASSIFIED

No.	Question	Expected answer	AI chatbot's answer	Accuracy
		<p>g. other occasions as ordered.</p>	<p>orders are not worn on lounge suits or blazers.</p> <p>3. When lounge suits, blazers and equivalent women's attire are worn in the evening in lieu of evening dress, it is quite acceptable to wear miniatures rather than full sized medals.</p> <p>4. Only those full-size insignia mounted on a medal bar may be worn on a civilian overcoat at outdoor functions. Neck badges are covered normally by outerwear. Stars of orders are not worn on overcoats.</p> <p>5. Included in the insignia of some orders and decorations is a lapel badge. This badge may be worn on the left lapel of a coat or jacket at any time when the full size or miniature insignia or the undress ribbons are not being worn. Women wear this badge in a similar position on a dress.</p> <p>6. When orders, decorations, and medals are worn, commendations and other authorized emblems shall not be obscured, and where necessary shall be repositioned lower, 2.6. cm from the bottom of the medals to the top edge of the senior commendation or emblem.</p> <p>(Learn more).</p>	
7	<p>Who can approve wearing a uniform in a theatrical production?</p>	<p>An officer commanding a command may authorize an officer or non-commissioned member to wear a uniform in a theatrical production, stage play or other public performance, where the officer is satisfied that no discredit to the service will ensue. (QR&O 17.065).</p> <p>1. by custom, veterans and other ex-service members may wear undress caps (e.g., berets, wedge caps), with badges, on remembrance and memorial occasions, subject to agreement</p>	<p>An officer commanding a command may authorize an officer or non-commissioned member to wear a uniform in a theatrical production, stage play or other public performance, where the officer is satisfied that no discredit to the service will ensue. (QR&O 17.065).</p> <p>1. by custom, veterans and other ex-service members may wear undress caps (e.g., berets, wedge caps), with badges, on remembrance and memorial</p>	100%

CAN UNCLASSIFIED

No.	Question	Expected answer	AI chatbot's answer	Accuracy
		<p>of the branch/regiment concerned;</p> <p>2. civilian band volunteers may be authorized to wear uniform as noted in Chapter 5, Section 2;</p> <p>3. members of Sea, Army and Air cadet organizations may wear CAF uniform items and affiliated unit insignia; and</p> <p>4. historical re-enactment groups may be authorized to wear obsolete uniforms subject to the agreement and general supervision of the branch/regiment or environment concerned.</p>	<p>occasions, subject to agreement of the branch/regiment concerned;</p> <p>2. civilian band volunteers may be authorized to wear uniform as noted in Chapter 5, Section 2;</p> <p>3. members of Sea, Army and Air cadet organizations may wear CAF uniform items and affiliated unit insignia; and</p> <p>4. historical re-enactment groups may be authorized to wear obsolete uniforms subject to the agreement and general supervision of the branch/regiment or environment concerned.</p> <p>(Learn more).</p>	
8	Am I required to purchase mess dress?	All Regular Force officers are required to be in possession of mess dress No. 2, which shall be procured at individual expense. Newly-commissioned officers are required to obtain this order of dress not later than six months after commissioning.	All Regular Force officers are required to be in possession of mess dress No. 2, which shall be procured at individual expense. Newly-commissioned officers are required to obtain this order of dress not later than six months after commissioning.	100%
9	Who can wear gorget patches?	General officers and Colonels will wear long gorget patches sewn onto the collar of the service dress jacket while in No. 1, No.1A and No.3 order of dress. Short gorget patches will be clipped onto the shirt collar while in No. 3B or 3C order of dress. Honourary appointments are to wear the unit or corps identifier on the collar in lieu of gorget patches.	General officers and Colonels will wear long gorget patches sewn onto the collar of the service dress jacket while in No. 1, No.1A and No.3 order of dress. Short gorget patches will be clipped onto the shirt collar while in No. 3B or 3C order of dress. Honourary appointments are to wear the unit or corps identifier on the collar in lieu of gorget patches.	100%

CAN UNCLASSIFIED

No.	Question	Expected answer	AI chatbot's answer	Accuracy
10	Where (or how) do I wear my rank on my turban?	<p>Sikh Turban. Rank embellishment shall be sewn centred on turban ribbons, when worn, as follows:</p> <ul style="list-style-type: none"> a. flag/general officers—to be promulgated; b. senior officers—gold nylon braid 2 cm wide; c. junior officers—gold nylon braid 0.6 cm wide; and d. officer/naval cadets—no embellishment. 	<p>Paratroop Personnel and Airborne Units. The maroon beret or turban shall be worn with all orders of dress, less orders of mess dress, by qualified parachutists wearing the Army uniform, when on the posted strength of an airborne unit, designated airborne sub-unit or element, and the Canadian Army Advanced Warfare Centre. The maroon beret or turban is further authorized for wear by qualified parachutists wearing the Army uniform when on staff, exchange, or liaison duties, and, while so employed, in receipt of paratroop allowance.</p> <p>(Learn more).</p>	0%
11	Can I wear my Flying badge with my cadpat or combat uniform?	<p>Air Force personnel shall wear the following headdress:</p> <ul style="list-style-type: none"> 1. ceremonial dress—wedge cap, toque, Yukon cap, or turban; 2. service dress—as above and beret; 3. operational dress: <ul style="list-style-type: none"> e. Flying clothing—wedge cap, beret, toque, or turban; and f. CADPAT—beret, toque, or turban. 	<p>Air Force personnel shall wear the following headdress:</p> <ul style="list-style-type: none"> 1. ceremonial dress—wedge cap, toque, Yukon cap, or turban; 2. service dress—as above and beret; 3. operational dress: <ul style="list-style-type: none"> 5. Flying clothing—wedge cap, beret, toque, or turban; and 6. CADPAT—beret, toque, or turban. <p>(Learn more).</p> <p><i>Note: Here the list numbers were changed from e. and f. to 5. and 6. Due to HTML rendering of a partial table.</i></p>	100%
12	What is a bronze, single oak leaf emblem?	<p>A bronze, single oak leaf emblem, normally worn on campaign ribbons (see sub-sub-paragraph 8.b. (4), above). If awarded under circumstances where no appropriate ribbon is available, it is worn as noted below.</p>	<p>A bronze, single oak leaf emblem, normally worn on campaign ribbons (see sub-sub-paragraph 8.b. (4), above). If awarded under circumstances where no appropriate ribbon is available, it is worn as noted below.</p> <p>(Learn more).</p>	100%

CAN UNCLASSIFIED

No.	Question	Expected answer	AI chatbot's answer	Accuracy
13	What are Canadian Ranger Authorized Items of dress?	<p>The CR uniform/orders of dress are a blend of authorized CR red items, CADPAT T/W and Army/CAF items:</p> <p>CR Authorized Items:</p> <ol style="list-style-type: none"> 1. Ball-cap: red with embroidered CR badge; 2. Sweatshirt: red hooded, with CR badge on front, rank tab centered above badge and Velcro removable Canadian Flag on left shoulder; 3. T-shirt: red CR badge on left chest and CAF flag on left sleeve; 4. CADPAT T/W pants; 5. CAF black belt; 6. Black WW boot; 7. Authorized Cold Weather Items: <ol style="list-style-type: none"> a. Red fleece toque; b. Red ICE jacket; c. Red fleece top or olive green thermal underwear top zipped up to neck are authorized as an undergarment only; 8. CADPAT T/W wind pants; and 9. Black/OG gauntlet gloves/glove CW. 	<p>The following paragraphs provide direction for Canadian Rangers (CR) on authorized variations of ceremonial, mess and service orders of dress; they do not provide direction on operational/field dress. See CR orders of dress Annex F. (Learn more)</p> <p><i>Note: Here the chatbot directs to the previous paragraph which is pointing to the expected answer</i></p>	95%
14	What non-operational headdress may I wear with flying clothing?	<p>The following non-operational headdress may be worn with flying clothing: wedge cap, beret, turban, balmoral, glengarry, caubeen, khaki tam-o'-shanter and toque (see Headdress Policy, Chapter 5, Section 1, paragraphs 2. to 6.).</p>	<p>The following non-operational headdress may be worn with flying clothing: wedge cap, beret, turban, balmoral, glengarry, caubeen, khaki tam-o'-shanter and toque (see Headdress Policy, Chapter 5, Section 1, paragraphs 2. to 6.) (Learn more).</p>	100%

CAN UNCLASSIFIED

No.	Question	Expected answer	AI chatbot's answer	Accuracy
15	When can I start wearing maternity dress?	When conventional pattern uniforms are no longer comfortable or aesthetically appropriate, a pregnant woman shall commence wearing maternity dress.	When conventional pattern uniforms are no longer comfortable or aesthetically appropriate, a pregnant woman shall commence wearing maternity dress. (Learn more)	100%
16	I just commissioned, how long do I have to obtain Mess Dress No. 2?	All Regular Force officers are required to be in possession of mess dress No. 2, which shall be procured at individual expense. Newly-commissioned officers are required to obtain this order of dress not later than six months after commissioning.	Acquisition 1. All Regular Force officers are required to be in possession of mess dress No. 2, which shall be procured at individual expense. Newly-commissioned officers are required to obtain this order of dress not later than six months after commissioning. 2. Mess dress No. 2 is optional for Regular Force non-commissioned members and all members of the Reserve Force. Acquisition is the responsibility of the individual. 3. Army colonels, on promotion to that rank, may elect to continue to wear their previous branch/regiment mess dress, with current rank insignia, instead of the pattern authorized for army colonels (less honorary and Royal appointees) and described in Chapter 5, Appendix 1 to Annex B, paragraph 9 . (Learn more) .	100%
17	What headdress do I wear with my Mess Dress?	Headdress is not worn with orders of mess dress	Headdress. Headdress is not worn with orders of mess dress. (Learn more) .	100%
18	I was issued a CD recently, how do I go about having it mounted?	The members have the option of choosing the means of having medals court mounted, e.g., getting the work done themselves and claiming reimbursement via a CF 52 General Allowance Claim or having the work done through base supply, who will	(1) Flying and specialist skill badges are not worn. (Learn more) .	0%

CAN UNCLASSIFIED

No.	Question	Expected answer	AI chatbot's answer	Accuracy
		<p>action via LPO or whatever means provides the best economies.</p> <ol style="list-style-type: none"> 1. Members claiming via CF 52 will be limited to amount determined by local commanders in accordance with financial regulations; 2. Miniature medals must be court-mounted at personal expense. 		
19	I was issued a CD recently, how do I get expenses paid to having it mounted?	<p>The members have the option of choosing the means of having medals court mounted, e.g., getting the work done themselves and claiming reimbursement via a CF 52 General Allowance Claim or having the work done through base supply, who will action via LPO or whatever means provides the best economies.</p> <ol style="list-style-type: none"> 1. Members claiming via CF 52 will be limited to amount determined by local commanders in accordance with financial regulations; 2. Miniature medals must be court-mounted at personal expense. 	<p>The members have the option of choosing the means of having medals court mounted, e.g., getting the work done themselves and claiming reimbursement via a CF 52 General Allowance Claim or having the work done through base supply, who will action via LPO or whatever means provides the best economies.</p> <ol style="list-style-type: none"> 1. Members claiming via CF 52 will be limited to amount determined by local commanders in accordance with financial regulations; 2. Miniature medals must be court-mounted at personal expense. <p>(Learn more).</p>	100%
20	Can I wear my oxford's or pumps on parade?	<p>Only ankle boots or shoes (oxfords) are authorized for wear on parade, except when wearing overalls. Wellington boots shall be worn with the latter. According to branch or regimental custom, officers and CWOs may wear Wellington boots with trousers.</p>	<p>Hose</p> <ol style="list-style-type: none"> 1. Nylons, Plain Pattern Skin-Toned <ol style="list-style-type: none"> 1. May be worn with oxfords or pumps, when a skirt is worn. 2. May be worn with oxfords or pumps, when slacks are worn. 2. Nylons, Plain Pattern Black. May be worn with oxfords or pumps, when slacks are worn, at no expense to the public. 3. Nylons, Plain Pattern, Ivory. May be worn with Navy summer service dress No. 3B, at no expense to the public. 	25%

CAN UNCLASSIFIED

No.	Question	Expected answer	AI chatbot's answer	Accuracy
			<p>(Learn more).</p> <p><i>Note: the following was given by the chatbot as the 4th best answer.</i></p> <p>Only ankle boots or shoes (oxfords) are authorized for wear on parade, except when wearing overalls. Wellington boots shall be worn with the latter. According to branch or regimental custom, officers and CWOs may wear Wellington boots with trousers.</p> <p>(Learn more).</p>	
21	Can I chew gum while in uniform ?	Personnel in uniform shall comport themselves in a manner which projects a positive military appearance. Behaviour such as chewing gum, slouching, placing hands in pockets, smoking or eating on the street and walking hand in hand, is forbidden. This instruction's objective is to project an image of a disciplined and self-controlled force.	Behaviour. Personnel in uniform shall comport themselves in a manner which projects a positive military appearance. Behaviour such as chewing gum, slouching, placing hands in pockets, smoking or eating on the street and walking hand in hand, is forbidden. This instruction's objective is to project an image of a disciplined and self-controlled force.	100%
22	Can I bring an umbrella to a parade?	umbrella (optional)—shall not be carried on parade	umbrella (optional)—shall not be carried on parade.	100%

CAN UNCLASSIFIED

No.	Question	Expected answer	AI chatbot's answer	Accuracy
23	Who can wear a pony tail?	<p>Women (see Figure 2-2-3). Hair shall not extend below the lower edge of the shirt collar. Exaggerated styles, including those with excessive fullness or extreme height, are not authorized. Braids, if worn, shall be styled conservatively and tied tightly: secured at the end by a knot or a small unadorned fastener. A second small unadorned fastener may be used to secure the top of the braid. A single braid shall be worn in the centre of the back. Double braids shall be worn behind the shoulders. Hair shall be a maximum length when gathered behind the head and braided which does not extend below the top of the armpit. Multiple braids (cornrows) shall be directed toward the back of the head, pulled tight to the head and secured at the end by a knot or a small-unadorned fastener. A second small unadorned fastener may be used to secure the top of the braid. Multiple braids extending below the lower edge of the collar are to be gathered in a bun. Pony tail shall be worn in the centre of the back. Hair shall be a maximum length when gathered behind the head and does not extend below the top of the armpit. With the permission of a Commanding Officer, a reasonable period may be authorized in order to transition from short to long hairstyles, during which time hair may extend below the lower edge of the shirt collar; all the while maintaining a positive military appearance, and subject to the member's safety.</p>	<p>Women (see Figure 2-2-3). Hair shall not extend below the lower edge of the shirt collar. Exaggerated styles, including those with excessive fullness or extreme height, are not authorized. Braids, if worn, shall be styled conservatively and tied tightly: secured at the end by a knot or a small unadorned fastener. A second small unadorned fastener may be used to secure the top of the braid. A single braid shall be worn in the centre of the back. Double braids shall be worn behind the shoulders. Hair shall be a maximum length when gathered behind the head and braided which does not extend below the top of the armpit. Multiple braids (cornrows) shall be directed toward the back of the head, pulled tight to the head and secured at the end by a knot or a small-unadorned fastener. A second small unadorned fastener may be used to secure the top of the braid. Multiple braids extending below the lower edge of the collar are to be gathered in a bun. Pony tail shall be worn in the centre of the back. Hair shall be a maximum length when gathered behind the head and does not extend below the top of the armpit. With the permission of a Commanding Officer, a reasonable period may be authorized in order to transition from short to long hairstyles, during which time hair may extend below the lower edge of the shirt collar; all the while maintaining a positive military appearance, and subject to the member's safety.</p> <p>(Learn more).</p>	100%
24	What is the length allowed for male hair?	<p>Hair shall be taper-trimmed at the back, sides, and above the ears to blend with the hair-style; be no more than 15 cm in length and sufficiently short that, when the hair is groomed and headdress is removed, no hair shall touch the ears or fall below the</p>	<p>The following additional details apply to specific groups to accord with religious and spiritual practices and public perceptions of a disciplined force:</p> <ol style="list-style-type: none"> 1. Men (see Figure 2-2-1). Hair shall be taper-trimmed at the back, 	100%

No.	Question	Expected answer	AI chatbot's answer	Accuracy
		<p>top of the eyebrows; be no more than 4 cm in bulk at the top of the head, gradually decreasing to blend with the taper-trimmed sides and back; and be kept free from the neck to a distance of 2.5 cm above the shirt collar. Taper trimmed square back styles and shaving of all the hair on the head are permitted.</p>	<p>sides, and above the ears to blend with the hair-style; be no more than 15 cm in length and sufficiently short that, when the hair is groomed and headdress is removed, no hair shall touch the ears or fall below the top of the eyebrows; be no more than 4 cm in bulk at the top of the head, gradually decreasing to blend with the taper-trimmed sides and back; and be kept free from the neck to a distance of 2.5 cm above the shirt collar. Taper trimmed square back styles and shaving of all the hair on the head are permitted....</p> <p>(Learn more).</p>	
25	Where do I put my poppy?	<p>The Royal Canadian Legion Poppy. The red poppy is an emblem of the RCL and is used to commemorate Canadians who died in battle. CAF members shall wear the poppy on all uniforms from the last Friday in October until Remembrance Day (November 11th), and are encouraged to wear it when attending any event whose main purpose is to commemorate Canadians who died in battle. Positioning of the poppy is illustrated in Figure 3-7-2 thru Figure 3-7-4.</p>	<p>The Royal Canadian Legion Poppy. The red poppy is an emblem of the RCL and is used to commemorate Canadians who died in battle. CAF members shall wear the poppy on all uniforms from the last Friday in October until Remembrance Day (November 11th), and are encouraged to wear it when attending any event whose main purpose is to commemorate Canadians who died in battle. Positioning of the poppy is illustrated in Figure 3-7-2 thru Figure 3-7-4.</p> <p>(Learn more).</p>	100%
26	Can the scarf be worn with windbreaker jacket?	<p>The scarf may be worn with the topcoat (gabardine), parka, and windbreaker jacket. https://www.canada.ca/en/service/s/defence/caf/military-identity-system/dress-manual/chapter-2/section-2.html.</p>	<p>Scarf. The scarf may be worn with the topcoat (gabardine), parka, and windbreaker jacket.</p> <p>(Learn more).</p>	100%

5 Future Work

While our initial results are intriguing there is still room for improvement for the question answering algorithm and the chatbot system.

Currently the question answering algorithm is retrieval based, meaning that all the responses it gives are directly from a policy. While this does ensure that all responses will be 100% accurate, some of the responses it returns are very long since they come from a policy document where length of text is not an issue. In order to solve this, we can use a language model such as GPT-3 [11] from OpenAI to perform text summarization on the response from the chatbot. Text summarization is an NLP task where a model attempts to shorten the input text as much as possible while maintaining all the important information. In doing so, the responses from the chatbot will be much shorter and easier to read for the users. Another advantage to using text summarization with GPT-3 is that it can be tailored to summarize to a specific user. Policy documents can be quite tricky to understand at times and use language that not people will understand. Using GPT-3 we can tell the model to summarize a policy body of text for a person in the general public. This step could also make the responses easier to read and more importantly easier to understand for the users.

The current version of the chatbot is also only trained to answer questions about the Dress Manual [9]. In the future, we would like to expand the proposed chatbot model to other policies and standards documents within the DND/CAF and Government of Canada.

6 Conclusion

In summary, given that policy documents can be long and difficult to understand, we developed an automated system of retrieving information from policy and standards documents using NLP. We explored the feasibility of using NLP to create a chatbot that can answer questions without the need for manual intervention. Our proposed chatbot algorithm used two BERT models in a retrieve and rank format. The first model used the bi-encoder uses BERT to retrieve the top 50 most relevant answers to the question using encoding and semantic search. The second model used the cross-encoder uses BERT to re-rank the retrieved answers based on the question and assigns each question-answer pair a score between 0 and 1 which signifies the confidence that the answer is the best.

We built a web application using Flask and React which allows users to interact with the chatbot and also provides feedback which can be used for model fine-tuning and future chatbot improvements. The web application can be run on a web browser without having the need to install as a separate application.

The evaluation of our model shows that the model performs quite well on basic questions however there is some room for improvement such as the need for follow up questions when the original question the user asked is not specific enough to retrieve the correct answer. Early testing shows about 85% accuracy in answering questions related to military dress policy; with further fine-tuning of the model with user feedback, we expect the accuracy to further improve in the long run.

References

- [1] National Defence, “Government of Canada,” Canada.ca, 01-Apr-2021. [Online]. Available: <https://www.canada.ca/en/department-national-defence/corporate/policies-standards.html>. [Accessed: 22-Jul-2021].
- [2] S. Dong, P. Wang and K. Abbas, “A survey on deep learning and its applications,” Computer Science Review, Volume 40, 2021, 100379, ISSN 1574-0137, <https://doi.org/10.1016/j.cosrev.2021.100379>. [Accessed: 10-Nov-2021].
- [3] C. Vu, “Advance NLP model via transferring knowledge from Cross-Encoders to Bi-Encoders,” Medium, 03-Jan-2021. [Online]. Available: <https://towardsdatascience.com/advance-nlp-model-via-transferring-knowledge-from-cross-encoders-to-bi-encoders-3e0fc564f554>. [Accessed: 08-Sep-2021].
- [4] J. Vig and K. Ramea. Comparison of transfer-learning approaches for response selection in multi-turn conversations. Workshop on DSTC7, 2019.
- [5] P. Nayak, “Understanding searches better than ever before,” Google Vice President, Search.[Online] Available: <https://blog.google/products/search/search-language-understanding-bert/>. [Accessed: 10-Nov-2021].
- [6] J. Devlin, M. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Cornell University, 2021.
- [7] A. Ravichander, A. Black, S. Wilson, T. Norton and N. Sadeh, “Question Answering for Privacy Policies: Combining Computational and Legal Perspectives,” ACL Anthology, 2021.
- [8] Y. Luan, J. Eisenstein, K. Toutanova and M. Collins, “Sparse, Dense, and Attentional Representations for Text Retrieval,” Cornell University, 2021.
- [9] National Defence, “Dress Instructions for the Canadian Forces - Canada.ca,” Canada.ca, 01-Apr-2021. [Online]. Available: <https://www.canada.ca/en/services/defence/caf/military-identity-system/dress-manual.html>. [Accessed: 22-Jul-2021].
- [10] Beautiful Soup Documentation, “Beautiful Soup Documentation - Beautiful Soup 4.9.0 documentation.” [Online]. Available: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>. [Accessed: 02-Sep-2021].
- [11] G. Brockman, “OpenAI API,” OpenAI, 21-Jun-2021. [Online]. Available: <https://openai.com/blog/openai-api/>. [Accessed: 08-Sep-2021].
- [12] <https://www.sqlite.org/index.html>. [Accessed: 10-Nov-2021].
- [13] A. Vaswani et al., “Attention Is All You Need,” Cornell University, 2021.

- [14] R. Horev, “BERT Explained: State of the art language model for NLP,” Medium, 2021. [Online]. Available: https://www.researchgate.net/figure/The-Transformer-based-BERT-base-architecture-with-twelve-encoder-blocks_fig2_349546860. [Accessed: 18-Nov-2021].
- [15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, “Attention Is All You Need,” Advances in neural information processing systems (pp. 5998–6008). [Online] <https://arxiv.org/abs/1706.03762>. [Accessed: 10-Nov-2021].
- [16] P. Joshi, “Transformers in NLP | State-Of-The-Art-Models,” Analytics Vidhya, 2019. [Online]. Available: <https://www.analyticsvidhya.com/blog/2019/06/understanding-transformers-nlp-state-of-the-art-models/>. [Accessed: 22-Jul-2021].
- [17] Softmax Function, “DeepAI,” 17-May-2019. [Online]. Available: <https://deepai.org/machine-learning-glossary-and-terms/softmax-layer>. [Accessed: 02-Sep-2021].
- [18] Y. Tamura, “Multi-head attention mechanism: ‘queries,’ ‘keys,’ and ‘values,’ over and over again—Data Science Blog,” Data-science-blog.com, 2021. [Online]. Available: <https://data-science-blog.com/blog/2021/04/07/multi-head-attention-mechanism/>. [Accessed: 22-Jul-2021].
- [19] TensorFlow Hub, “Tfhub.dev,” 2021. [Online]. Available: https://tfhub.dev/tensorflow/small_bert/bert_en_uncased_L-2_H-128_A-2/1. [Accessed: 22-Jul-2021].
- [20] “Okapi BM25—Wikipedia,” En.wikipedia.org, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Okapi_BM25. [Accessed: 22-Jul-2021].
- [21] “Pretrained Models—Sentence-Transformers documentation,” Sbert.net, 2021. [Online]. Available: https://www.sbert.net/docs/pretrained_models.html. [Accessed: 22-Jul-2021].
- [22] “MS MARCO (Microsoft Machine Reading Comprehension),” GitHub, 2021. [Online]. Available: <https://github.com/microsoft/MSMARCO-Passage-Ranking>. [Accessed: 22-Jul-2021].
- [23] S. Humeau, K. Shuster, M. Lachaux and J. Weston, “Poly-encoders: architectures and pre-training strategies for fast and accurate multi-sentence scoring,” Cornell University, 2020.
- [24] “Multilingual Sentence & Image Embeddings with BERT,” GitHub, 2021. [Online]. Available: <https://github.com/UKPLab/sentence-transformers>. [Accessed: 22-Jul-2021].
- [25] “cross-encoder/ms-marco-TinyBERT-L-4 Hugging Face,” Huggingface.co, 2021. [Online]. Available: <https://huggingface.co/cross-encoder/ms-marco-TinyBERT-L-4>. [Accessed: 22-Jul-2021].
- [26] “React – A JavaScript library for building user interfaces,” A JavaScript library for building user interfaces. [Online]. Available: <https://reactjs.org/>. [Accessed: 02-Sep-2021].
- [27] “Welcome to Flask—Flask Documentation (2.0.x),” Flask.palletsprojects.com, 2021. [Online]. Available: <https://flask.palletsprojects.com/en/2.0.x/>. [Accessed: 22-Jul-2021].

- [28] “Microsoft Bot Framework,” Dev.botframework.com, 2021. [Online]. Available: <https://dev.botframework.com>. [Accessed: 22-Jul-2021].
- [29] Docker Documentation, 27-Aug-2021. [Online]. Available: <https://docs.docker.com/>. [Accessed: 02-Sep-2021].
- [30] <https://www.sqlite.org/index.html>. [Accessed 09-Aug-2021].
- [31] C. Gunasekara and N. Chalifour, “Military Policy and Standards Question-Answering Chatbot Demo,” Presented to CWOs. D. Mistry and J. Proulx, 17-Aug-2021.
- [32] C. Gunasekara, N. Chalifour and M. Triff “Military Policy and Standards Question-Answering Chatbot Demo,” Presented to CWO. M.J.H. Johansen, 27-July-2021.
- [33] C. Gunasekara, N. Chalifour and M. Triff “Military Policy and Standards Question-Answering Chatbot Demo,” Presented to CPO J.M. Gagnon 31-March-2021.
- [34] “Fine-tuning a pretrained model,” Fine-tuning a pretrained model - transformers 4.10.0 documentation. [Online]. Available: <https://huggingface.co/transformers/training.html>. [Accessed: 08-Sep-2021].

List of Symbols/Abbreviations/Acronyms/Initialisms

ADM(DIA)	Assistant Deputy Minister (Data, Innovation, Analytics)
AI	artificial intelligence
API	application programming interface
BERT	Bidirectional Encoder Representations from Transformers
CAF	Canadian Armed Forces
CORA	Centre for Operational Research and Analysis
DND	Department of National Defence
DRDC	Defence Research and Development Canada
GPT	Generative Pre-trained Transformer
HTML	HyperText Markup Language
IR	information retrieval
NLP	natural language processing
PaaS	platform-as-a-service
RNN	Recurrent Neural Networks
UI	user interface
VM	virtual machine

CAN UNCLASSIFIED

DOCUMENT CONTROL DATA		
*Security markings for the title, authors, abstract and keywords must be entered when the document is sensitive		
1. ORIGINATOR (Name and address of the organization preparing the document. A DRDC Centre sponsoring a contractor's report, or tasking agency, is entered in Section 8.) DRDC – Centre for Operational Research and Analysis Defence Research and Development Canada Carling Campus, 60 Moodie Drive, Building 7S.2 Ottawa, Ontario K1A 0K2 Canada	2a. SECURITY MARKING (Overall security marking of the document including special supplemental markings if applicable.) CAN UNCLASSIFIED	
	2b. CONTROLLED GOODS NON-CONTROLLED GOODS DMC A	
3. TITLE (The document title and sub-title as indicated on the title page.) Question Answering Artificial Intelligence Chatbot on Military Dress Policy: A Natural Language Processing Based Application		
4. AUTHORS (Last name, followed by initials – ranks, titles, etc., not to be used) Gunasekara, C.; Chalifour, N.; Triff, M.		
5. DATE OF PUBLICATION (Month and year of publication of document.) November 2021	6a. NO. OF PAGES (Total pages, including Annexes, excluding DCD, covering and verso pages.) 38	6b. NO. OF REFS (Total references cited.) 34
7. DOCUMENT CATEGORY (e.g., Scientific Report, Contract Report, Scientific Letter.) Scientific Report		
8. SPONSORING CENTRE (The name and address of the department project office or laboratory sponsoring the research and development.) DRDC – Centre for Operational Research and Analysis Defence Research and Development Canada Carling Campus, 60 Moodie Drive, Building 7S.2 Ottawa, Ontario K1A 0K2 Canada		
9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.) ADM(DIA)_001	9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)	
10a. DRDC PUBLICATION NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.) DRDC-RDDC-2021-R175	10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)	
11a. FUTURE DISTRIBUTION WITHIN CANADA (Approval for further dissemination of the document. Security classification must also be considered.) Public release		
11b. FUTURE DISTRIBUTION OUTSIDE CANADA (Approval for further dissemination of the document. Security classification must also be considered.)		
12. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Use semi-colon as a delimiter.) natural language processing; artificial intelligence		

13. ABSTRACT (When available in the document, the French version of the abstract must be included here.)

In the Canadian Armed Forces (CAF) there are many policies and standards that must be followed. These policy documents can be quite long and sometimes difficult to navigate. For example, the Department of National Defence (DND)/CAF corporate policies in [1] are available in nested weblinks with seemingly overlapping titles with various chapters; each comprises hundreds of paragraphs with no direct search option strictly for the policy document. In a situation where someone needs information that may be in a policy, it can take quite a bit of time to get an answer when scrolling through a very comprehensive policy or standard document. This Scientific Report explores how artificial intelligence (AI) using natural language processing (NLP) can be used to develop a question answering chatbot that can retrieve passages from a policy to answer a user's question.

We first develop a data collection method to automatically read the policy and standards documents and then implement an NLP based data retrieval system to find the most suitable answer to users' questions. We also introduce a mechanism for the chatbot to self-learn using the feedback sent by the users, allowing it to adaptively change its algorithm to improve the accuracy. The chatbot was tested against sample questions sent by CAF members and early testing shows a high accuracy in the responses.

Dans les Forces armées canadiennes (FAC), de nombreuses politiques et normes doivent être respectées. Les documents énonçant ces politiques et normes peuvent se révéler très volumineux, et il est parfois difficile de s'y retrouver. Par exemple, les politiques ministérielles du ministère de la Défense nationale (MDN)/CAF dans [1] sont accessibles sous forme de liens Internet imbriqués avec des titres qui semblent se chevaucher. Ils comprennent plusieurs chapitres et chaque chapitre contient des centaines de paragraphes. Il n'existe aucune fonction permettant d'effectuer une recherche directe ciblée dans ledit document. Par conséquent, lorsqu'une personne doit rechercher un renseignement, il lui faut parfois beaucoup de temps pour parcourir ce document exhaustif avant d'obtenir une réponse. Ce rapport scientifique explore comment l'intelligence artificielle (IA) utilisant le traitement du langage naturel peut être utilisée pour développer un agent conversationnel qui peut trouver les passages pertinents dans une politique pour répondre à la question d'un utilisateur.

Nous développons tout d'abord une méthode de collecte de données pour lire automatiquement les documents relatifs aux politiques et aux normes, puis nous mettons en œuvre un système d'extraction de données basé sur le traitement du langage naturel pour trouver la réponse la plus appropriée aux questions des utilisateurs. Nous introduisons également un mécanisme permettant à l'agent conversationnel d'apprendre par lui-même en utilisant le retour d'information envoyé par les utilisateurs, ce qui lui permet de modifier son algorithme de manière adaptative pour améliorer la précision. L'agent conversationnel a été testé sur des exemples de questions envoyées par les membres des FAC et les premiers tests démontrent une grande précision dans les réponses.