Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-021-03439-8

ORIGINAL RESEARCH

Chatbot Interaction with Artificial Intelligence: human data augmentation with T5 and language transformer ensemble for text classification

Jordan J. Bird[1] · **Anikó Ekárt**[2] · **Diego R. Faria**[1]

Received: 21 August 2020 / Accepted: 5 August 2021

Abstract
In this work we present the Chatbot Interaction with Artificial Intelligence (CI-AI) framework as an approach to the training of a transformer based chatbot-like architecture for task classification with a focus on natural human interaction with a machine as opposed to interfaces, code, or formal commands. The intelligent system augments human-sourced data via artificial paraphrasing in order to generate a large set of training data for further classical, attention, and language transformation-based learning approaches for Natural Language Processing (NLP). Human beings are asked to paraphrase commands and questions for task identification for further execution of algorithms as skills. The commands and questions are split into training and validation sets. A total of 483 responses were recorded. Secondly, the training set is paraphrased by the T5 model in order to augment it with further data. Seven state-of-the-art transformer-based text classification algorithms (BERT, DistilBERT, RoBERTa, DistilRoBERTa, XLM, XLM-RoBERTa, and XLNet) are benchmarked for both sets after fine-tuning on the training data for two epochs. We find that all models are improved when training data is augmented by the T5 model, with an average increase of classification accuracy by 4.01%. The best result was the RoBERTa model trained on T5 augmented data which achieved 98.96% classification accuracy. Finally, we found that an ensemble of the five best-performing transformer models via Logistic Regression of output label predictions led to an accuracy of 99.59% on the dataset of human responses. A highly-performing model allows the intelligent system to interpret human commands at the social-interaction level through a chatbot-like interface (e.g. "Robot, can we have a conversation?") and allows for better accessibility to AI by non-technical users.

Keywords Chatbot · Human-machine interaction · Data augmentation · Transformers · Language transformation · Natural Language Processing

1 Introduction

Attention-based and transformer language models are a rapidly growing field of study within machine learning and artificial intelligence and for applications beyond. The field of Natural Language Processing (NLP) has especially been advanced through transformers due to their approach to reading being more akin to human behaviour than classical sequential techniques. With many industries turning to Artificial Intelligence (AI) solutions by the day, models have a growing requirement for robustness, explainability, and accessibility since AI solutions are becoming more and more popular for those without specific technical backgrounds in the field. Another interesting field that is similarly being seen more often is that of data augmentation; that is, creating data from a set that in itself

✉ Jordan J. Bird
birdj1@aston.ac.uk

Anikó Ekárt
a.ekart@aston.ac.uk

Diego R. Faria
d.faria@aston.ac.uk

[1] Aston Robotics, Vision and Intelligent Systems Lab (ARVIS Lab), Aston University, Birmingham, UK

[2] School of Engineering and Applied Science, Aston University, Birmingham, UK

Published online: 23 August 2021

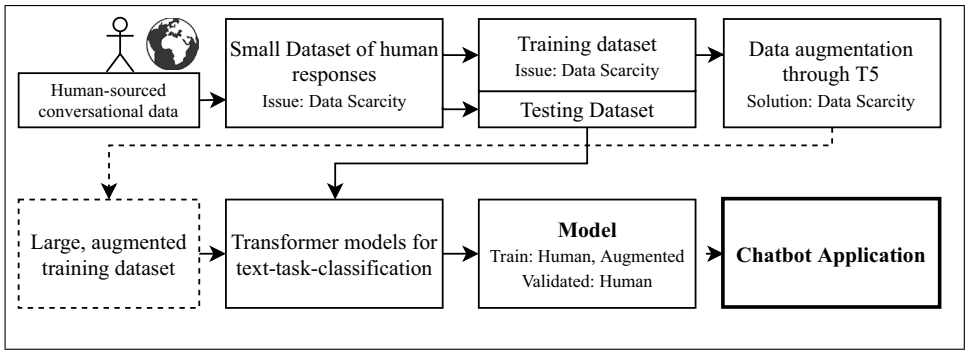

Fig. 1 A general overview of the proposed approach

increases the quality of that set of data. The alternative to data augmentation, which is unfortunately the case with many modern NLP systems, is to gather more data. As an alternative to unwanted privacy concerns, data scientists may instead find ways to augment the data as a friendlier alternative.

In this study, we bring together all of these aforementioned concepts and fields of study to form a system that we call Chatbot Interaction with Artificial Intelligence (CI-AI). A general overview of the approach can be observed in Fig. 1. As an alternative to writing code and managing data, complex machine learning tasks such as conversational AI, sentiment analysis, scene recognition, brainwave classification and sign language recognition among others are given accessibility through an interface of natural, social interaction via both verbal and non-verbal communication. That is, for example, a spoken command of "can we have a conversation?" or a sign language command of "can-we-talk" would command the system to launch a conversational AI program. For such a system to be possible, it needs to be robust, since an interactive system that makes one mistake for many successes would be considered a broken system. The system needs to be accessible to a great number of people with differing backgrounds, and thus must have the ability to generalise by being exposed to a large amount of training data. Last, but by no means least, the system needs to be explainable; as given in a later example, if a human were to utter the phrase, "Feeling sad today. Can you cheer me up with a joke?", which features within that phrase lead to a correct classification and command to the chatbot to tell a joke? Where does the model focus within the given text in order to correctly predict and fulfil the human's request? Thus, to achieve these goals, the scientific contributions of this work are as follows:

1. The collection of a seven-class command-to-task dataset from multiple human beings from around the world, giving a total of 483 data objects.
2. Augmentation of the human data with a transformer-based paraphrasing model which results in a final training dataset of 13,090 labelled data objects.
3. Benchmarking of seven State-of-the-Art transformer-based classification approaches for text-to-task commands. Each model is trained on the real training data and validation data, and is then trained on the real training data plus the paraphrased augmented data and validation data. We find that all seven models are improved significantly when exposed to augmented data.
4. A deep exploration of the best model. Firstly in order to discern the small amount of errors (1.04% errors) and how they were caused by seeing the largest errors in terms of loss and the class probability distributions. Secondly, the chatbot is given commands that were not present during training or validation, and top features (words) are observed- interestingly, given their technical nature, the models focus keenly on varying parts of the sentence similar to a human reading.
5. Stacked Generalisation approaches are explored in order to ensemble several highly performing models, results show that the stack of multiple transformers outperform the best singular model.

The rest of this article is structured as follows. Initially, the background and related studies are explored in Sect. 2. The method of the experiments are described in Sect. 3, and the results from the experiments are then presented in Sect. 4. With the best-performing model in mind, Sect. 4.1 then explores the model in terms of the small number of errors made, and how the model interprets new and unseen data (ie. should the model be in deployment). Finally, conclusions are drawn and future work is suggested in Sect. 5.

2 Background and related works

Data scarcity often poses a problem in the field of NLP (Roller et al. 2020), given that even a large subject set of over one hundred individuals may still result in a relatively small amount of data collected in comparison to other fields, with consideration to the size of data usually required for machine and deep learning models. Several works have suggested that data augmentation is an important solution

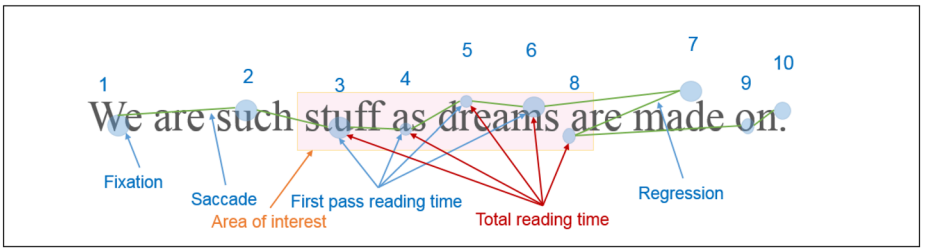

Fig. 2 An eye-tracking study of natural reading from (Eckstein et al. 2019). The reader's gaze naturally follows a left-to-right reading pattern with a fluctuation back to the main area of interest, where the main reading time is greater than that of the rest of the sentence

to these problems, that is, engineering synthetic data to increase the size of a dataset. It is important that the synthetic data is not only different to the actual data, but also that it contains useful knowledge to improve classifiers when attempting to understand language. For example, chatbot software has been noted to improve in ability when synonymous terms are generalised as flags (Bird et al. 2018a). Techniques that have shown promise include random token perturbations (Wei and Zou 2019), back-translation (Shleifer 2019), and inductive transfer learning (Howard and Ruder 2018). Recently, it was noted that paraphrasing provides a strong candidate for solving data scarce NLP problems (Bannard and Callison-Burch 2005; Marton et al. 2009; Lewis et al. 2020) as well as language transformation (Sun et al. 2020). In this work, we consider improving a data scarce problem by augmenting the training dataset by paraphrasing it via a pre-trained Transformer model. In addition, the text classification models themselves are also transformative in nature.

The Transformer is a new concept in the field of deep learning (Vaswani et al. 2017). Transformers currently have a primary focus on NLP, but state-of-the-art image processing using similar networks have recently been explored (Qi et al. 2020). With the idea of *paying attention* in mind, the theory behind the exploration of Transformers in NLP is their more natural approach to sentences; rather than focusing on one token at a time in the order that they appear and suffering from the vanishing gradient problem (Schmidhuber 1992), Transformer-based models instead pay attention to tokens in a learned order and as such enable more parallelisation while improving upon many NLP problems through which many benchmarks have been broken (Vaswani et al. 2017; Wang et al. 2018). For these reasons, such approaches are rapidly forming State-of-the Art scores for many NLP problems (Tenney et al. 2019). For text data in particular these include generation (Devlin and Chang 2018; Radford et al. 2019), question answering (Shao et al. 2019; Lukovnikov et al. 2019), sentiment analysis (Naseem et al. 2020; Shangipour ataei et al. 2020), translation (Zhang et al. 2018; Wang et al. 2019b; Di Gangi et al. 2019), paraphrasing (Chada 2020; Lewis et al. 2020), and classification (Sun et al. 2019; Chang et al. 2019). According to (Vaswani et al. 2017), Transformers are based on calculation of scaled dot-product attention units. These weights are calculated for each word within the input vector of words (document or sentence). The output of the attention unit are embeddings for a combination of relevant tokens within the input sequence. This is shown later on in Sect. 4.1 where both correctly and incorrectly classified input sequences are highlighted with top features that lead to such a prediction. Weights for the query W_q, key W_k, and value W_v are calculated as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{1}$$

The query is an object within the sequence, the keys are vector representations of said input sequence, and the values are produced given the query against keys. Unsupervised models receive Q, K and V from the same source and thus pay *self-attention*. For tasks such as classification and translation, K and V are derived from the source and Q is derived from the target. For example, Q could be a class for the text to belong to ie. for sentiment analysis "positive" and "neutral" and thus the prediction of the classification model. Secondly, for translation, values K and V could be derived from the English sentence "Hello, how are you?" and Q the sequence "Hola, como estas?" for supervised English-Spanish machine translation. All of the State-of-the-Art models benchmarked in these experiments follow the concept of Multi-headed Attention. This is simply a concatenation of multiple i attention heads h_i to form a larger network of interconnected attention units:

$$\begin{aligned} MultiHead(Q, K, V) &= Concatenate(head_1, ..., head_h)W^O \\ head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V). \end{aligned} \tag{2}$$

It is important to note that human beings also do not read in a token-sequential nature as is with classical models such as the Long Short Term Memory (LSTM) network (Hochreiter and Schmidhuber 1997). Figure 2 from a 2019 study on reading comprehension (Eckstein et al. 2019) shows human behaviour while reading. It can be observed from this example and other related studies (Shagass et al. 1976; Kruger and Steyn 2014; Wang et al. 2019a), that rather than simply reading left-to-right (or right-to-left (Wang et al. 2019a; Marquis et al. 2020)), instead attention is paid to areas of

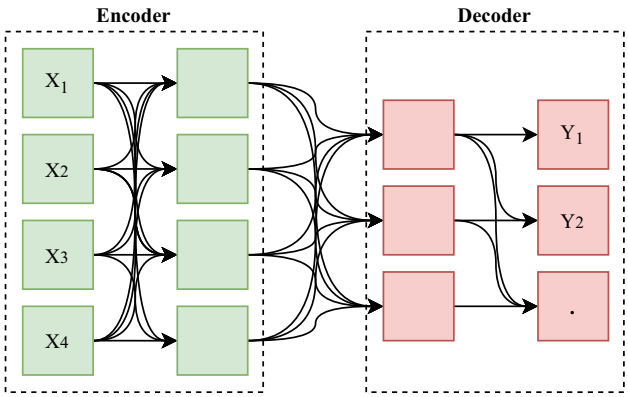

Fig. 3 Diagram of an encoder–decoder architecture

interest within the document. Of course, a human being does not follow the equations previously described, but it can be noted that attention-based models are more similar to human reading comprehension than that of sequential models such as the LSTM. Later, in Sect. 4.1, during the exploration of top features within correct classifications, it can be observed that RoBERTa also focuses upon select areas of interest within a text for prediction.

The Text-to-Text Transfer Transformer (T5) model is a unified approach to text transformers from Google AI (Raffel et al. 2019). T5 aims to unify NLP tasks by restricting output to text which is then interpreted to score the learning task; for example, it is natural to have a text output for a translation task (as per the previous example on English-Spanish translation), but for classification tasks on the other hand, a sparse vector for each prediction is often expected—T5 instead would output a textual representation of the class(es). This feature allows T5 to be extended to many NLP tasks outside of those suggested and benchmarked in the original work. To give a specific example to this study, an English–English translation of example "what time is it right now?" to "could you tell me the time, please?" provides a paraphrasing activity. That is, to express the same meaning of a text written in a different way. *Text-to-text* formatted problems such as paraphrasing are enabled due to T5's encoder–decoder architecture, a diagram of which can be observed in Fig. 3. The model is trained via teacher forcing (Williams and Zipser 1989; Goodfellow et al. 2017) where ground truth is used as input; each training instance requires a target for each input sequence. For example in sequence-to-sequence, an output with an early mistake in the sequence would be punished for every subsequent output, whereas teacher-forcing allows for the discarding of early mistakes after calculating the error at that step. Ultimately this leads to a learning process wherein statistical properties can be calculated quicker. Each encoder and decoder performs self attention and encoder–decoder attention as can be observed in Eq. 1.[1]

Chatbots are a method of human-machine interaction that have transcended novelty to become a useful technology of the modern world. A biological signal study from 2019 (Muscular activity, respiration, heart rate, and electrical behaviours of the skin) found that textual chatbots provide a more comfortable platform of interaction than with more human-like animated avatars, which caused participants to grow uncomfortable within the uncanny valley (Ciechanowski et al. 2019). Many chatbots exist as entertainment and as forms of art, such as in 2018 (Candello et al. 2018) when natural interaction was enabled via state-of-art of the art methods for character generation from text (Haller and Rebedea 2013). This allowed for 10,000 visitors to converse with 19th century characters from Machado de Assis' "Dom Casmurro". It has been strongly suggested through multiple experiments that natural interaction with chatbots will provide a useful educational tool in the future for students of varying ages (Kerlyl et al. 2006; Leonhardt et al. 2007; Bollweg et al. 2018). The main open issue in the field of conversational agents is data scarcity which in turn can lead to unrealistic and unnatural interaction, overcoming which are requirements for the Loebner Prize based on the Turing test (Stephens 2002). Solutions have been offered such as data selection of input (Dimovski et al. 2018), input simplification and generalisation (Bird et al. 2018a), and more recently paraphrasing of data (Virkar et al. 2019). These recent advances in data augmentation by paraphrasing in particular have shown promise in improving conversational systems by increasing understanding of naturally spoken language (Hou et al. 2018; Jin et al. 2018).

3 Proposed approach

In this section, the proposed approach followed by the experiments are described, from data collection to modes of learning and classification. The main aim of this work is to enable accessibility to previous studies, and in particular the machine learning models derived throughout them. Accessibility is presented in the form of social interaction, where a user requests to use a system in particular via natural language and the task is derived and performed. The seven commands are:

– Scene Recognition (Bird et al. 2020b)—The participant requests a scene recognition algorithm to be instantiated, a camera and microphone are activated for multi-modality classification.

[1] Further detail on T5 can be found in (Raffel et al. 2019).

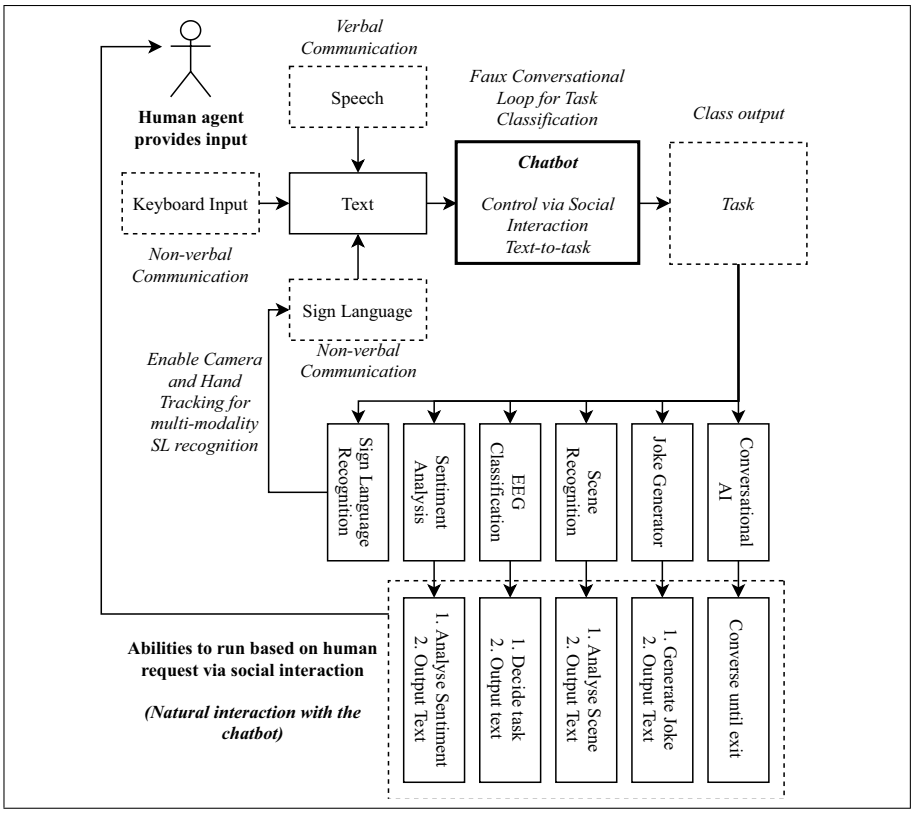

Fig. 4 Overall view of the Chatbot Interaction with Artificial Intelligence (CI-AI) system as a looped process guided by human input, through natural social interaction due to the language transformer approach. The chatbot itself is trained via the process in Fig. 5

- EEG Classification—The participant requests an EEG classification algorithm to be instantiated and begins streaming data from a MUSE EEG headband, there are two algorithms:
  - EEG Mental State Classification (Bird et al. 2018b)—Classification of whether the participant is concentrating, relaxed, or neutral.
  - EEG Emotional State Classification (Bird et al. 2019a)—Classification of emotional valence, positive, negative, or neutral.
- Sentiment Analysis of Text (Bird et al. 2019b)—The participant requests to instantiate a sentiment analysis classification algorithm for a given text.
- Sign Language Recognition (Bird et al. 2020a)—The participant requests to converse via sign language, a camera and Leap Motion and Leap Motion are activated for multi-modality classification. Sign language is now accepted as input to the task-classification layer of the chatbot.
- Conversational AI (Bird et al. 2018a)—The participant requests to have a conversation, a chatbot program is executed.
- Joke Generator (Manurung et al. 2008; Petrović and Matthews 2013)—The participant requests to hear a joke, a joke-generator algorithm is executed and output is printed.

Each of the given commands are requested in the form of natural social interaction (either by keyboard input, speech converted to text, or sign language converted to text), and through accurate recognition, the correct algorithm is executed based on classification of the human input. Tasks such as sentiment analysis of text and emotional recognition of EEG brainwaves, and mental state recognition compared to emotional state recognition, are requested in similar ways and as such constitutes a difficult classification problem. For these problems, minute lingual details must be recognised in order to overcome ambiguity within informal communication.

Figure 4 shows the overall view of the system. Keyboard input text, or speech and sign language converted to text provide an input of natural social interaction. The chatbot, trained on the tasks, classifies which task has been requested and executes said task for the human participant. Sign language, due to its need for an active camera and hand-tracking, is requested and activated via keyboard input or speech and itself constitutes a task. In order to derive the bold 'Chatbot' module in Fig. 5 shows the training processes followed. Human data is gathered via questionnaires which gives a relatively small dataset

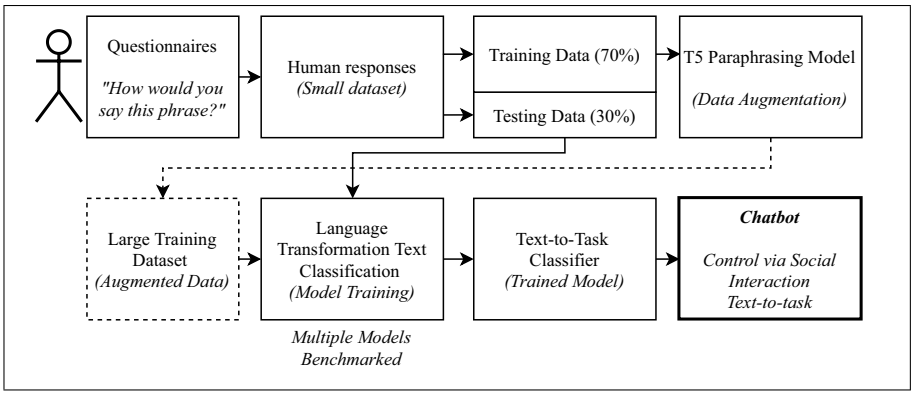

Fig. 5 Data collection and model training process. In this example, the T5 paraphrasing model is used to augment and enhance the training dataset. Models are compared when they are augmented and when they are not on the same validation set, in order to discern what affect augmentation has

Table 1 A selection of example statements presented to the users for paraphrasing

Example Statement	Class
"Would you like to talk?"	CHAT
"Tell me a joke"	JOKE
"Can you tell what mood I'm in from my brainwaves?"	EEG-EMOTIONS
"Am I concentrating? Or am I relaxed?	EEG-MENTAL-STATE
"Look around and tell me where you are"	SCENE-CLASSIFICATION
"Is this message being sarcastic or are they genuine?"	SENTIMENT-ANALYSIS
"I cannot hear the audio, please sign instead"	SIGN-LANGUAGE

One example is given for each for readability purposes, but a total of five examples were presented to the participants

(even though many responses were gathered, the nature of NLP tends to require a large amount of mined data), split into training and testing instances. The first experiment is built upon this data, and State-of-the-Art transformer classification models are benchmarked. In the second set of more complex experiments, the T5 paraphrasing model augments the training data and generates a large dataset, which are then also benchmarked with the same models and validation data in order to provide a direct comparison of the effects of augmentation. Augmentation is performed by paraphrasing the data within the training set, which therefore provides a greater number of training examples. Several metrics are used to compare models in terms of True Positives (*TP*), True Negatives (*TN*), False Positives (*FP*), and False Negatives (*FN*).

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}. \tag{3}$$

Precision:

$$Precision = \frac{TP}{TP + FP}. \tag{4}$$

Recall:

$$Recall = \frac{TP}{TP + FN}. \tag{5}$$

And finally the F1-Score:

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{6}$$

A questionnaire was published online for users to provide human data in the form of examples of commands that would lead to a given task classification. Five examples were given for each, and Table 1 shows some examples that were presented. The questionnaire instructions were introduced with "For each of these questions, please write how you would state the text differently to how the example is given. That is, paraphrase it. Please give only one answer for each. You can be as creative as you want!". Two examples were given that were not part of any gathered classes, "If the question was: 'How are you getting to the cinema?' You could answer: 'Are we driving to the cinema or are we getting the bus?' and "If the question was: 'What time is it?' You could answer: 'Oh no, I slept in too late... Is it the morning or afternoon? What's the time?"'. These examples were designed to show the users that creativity and diversion from the given example was not just acceptable but also

Table 2 An overview of models benchmarked and their topologies

Model	Topology
BERT (Devlin et al. 2018)	12-layer, 768-hidden, 12-heads, 110 M parameters
DistilBERT (Sanh et al. 2019)	6-layer, 768-hidden, 2-heads, 66 M parameters
RoBERTa (Liu et al. 2019)	12-layer, 768-hidden, 12-heads, 125 M parameters
DistilRoBERTa (Liu et al. 2019; Wolf et al. 2019)	6-layer, 768-hidden, 12-heads, 82 M parameters
XLM (Conneau and Lample 2019)	12-layer, 2048-hidden, 16-heads, 342 M parameters
XLM-RoBERTa (Conneau et al. 2019)	12-layer, 768-hidden, 3072 feed-forward, 8-heads, 125 M parameters
XLNet (Yang et al. 2019)	12-layer, 768-hidden, 12-heads, 110 M parameters
T5 (Paraphraser) (Raffel et al. 2019)	12-layer, 768-hidden, 12-heads, 220 M parameters

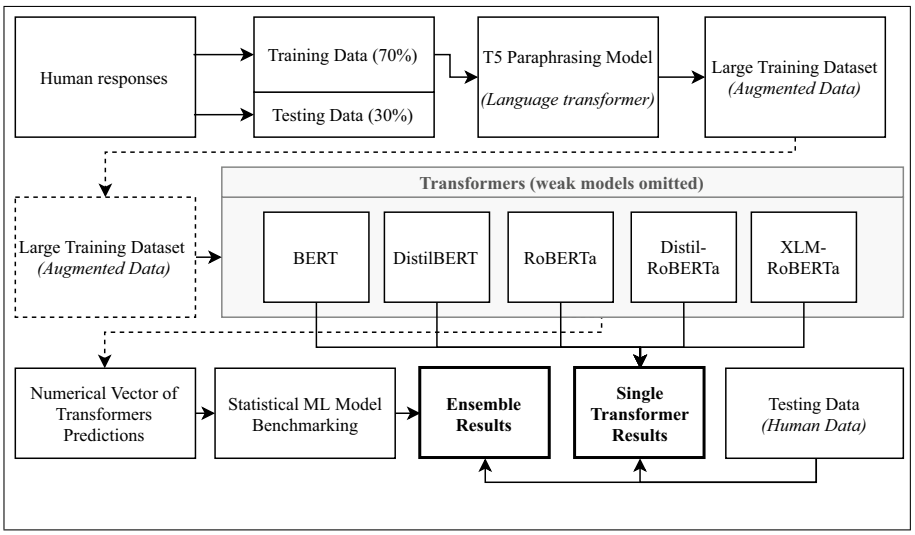

Fig. 6 An ensemble strategy where statistical machine learning models trained on the predictions of the transformers then classify the text based on the test data predictions of the transformer classification models

encouraged, so long as the general meaning and instruction of and within the message was retained (the instructions ended with "The example you give must still make sense, leading to the same outcome."). Extra instructions were given as and when requested, and participants did not submit any example phrases nor were any duplicates submitted. A total of 483 individual responses were recorded. The answers were split 70/30 on a per-class basis to provide two class-balanced datasets, firstly for training (and augmentation), and secondly for validation. That is, regardless of augmentation, the model is tested based on this validation set and are all thus directly comparable in terms of their learning abilities. The T5 paraphrasing model which was trained on the Quora question pairs dataset (Quora 2017) is executed a maximum of 50 times for each statement within the training set, where the model will stop generating paraphrases if the limit of possibilities or 50 total are reached. Once each statement had been paraphrased, a random subsample of the dataset on a per-class basis was taken set at the number of data objects within the least common class (sign language). Concatenated then with the real training data, a dataset of 13,090 examples were formed (1870 per class). This dataset thus constitutes the second training set for the second experiment, in order to compare the effects of data augmentation for the problem presented. The datasets for these experiments are publicly available.[2]

Table 2 shows the models that are trained and benchmarked on the two training sets (Human, Human+T5), and validated on the same validation dataset. It can be observed that the models are complex, and training requires a relatively high amount of computational resources. Due to this, the pre-trained weights for each model are fine-tuned for two epochs on each of the training datasets.

3.1 Statistical ensemble of transformer classifiers

Finally, a further experiment is devised to combine the results of the best models within an ensemble, which can be observed in Fig. 6. The training and test datasets

[2] https://www.kaggle.com/birdy654/human-robot-interaction-via-t5-data-augmentation.

Table 3 Classification results of each model on the same validation set, both with and without augmented paraphrased data within the training dataset

Model	With T5 paraphrasing				Without T5 paraphrasing			
	Acc.	*Prec.*	*Rec.*	*F1*	*Acc.*	*Prec.*	*Rec.*	*F1*
BERT	**98.55**	0.99	0.99	0.99	90.25	0.93	0.9	0.9
DistilBERT	**98.34**	0.98	0.98	0.98	97.3	0.97	0.97	0.97
DistilRoBERTa	**98.55**	0.99	0.99	0.99	95.44	0.96	0.95	0.95
RoBERTa	**98.96**	0.99	0.99	0.99	97.93	0.98	0.98	0.98
XLM	**14.81**	0.15	0.15	0.15	13.69	0.02	0.14	0.03
XLM-RoBERTa	**98.76**	0.99	0.99	0.99	87.97	0.9	0.88	0.88
XLNet	**35.68**	0.36	0.35	0.36	32.99	0.33	0.24	0.24
Average	77.66	0.78	0.78	0.78	73.65	0.73	0.72	0.71

Bold shows best model per run, underline shows the best model overall

are firstly distilled into a numerical vector of five predictions made by the five selected transformer models. These features are analysed in terms of classification ability by way of their relative entropy. That is the change in entropy ($E(s) = -\sum_j p_j \times log(p_j)$) in terms of the classification of a set P_j with solution s. Relative entropy or information gain is thus given as $InfoGain(T, a) = E(T) - E(T|a)$ in regards to the calculated Entropy E, for instances of original ruleset $H(T)$ and comparative ruleset $H(T|a)$. Following this, statistical machine learning models are trained on the training set and validated by the test set in order to discern whether combining the models together ultimately improves the ability of the model. The reasoning behind a statistical ensemble is that it enables possible improvements to a decision system's robustness and accuracy (Zhang and Ma 2012). Given that nuanced differences between the transformers may lead to 'personal' improvements in some situations and negative impacts in others, for example when certain phrases appear within commands, a more democratic approach may allow the pros of some models outweigh the cons of others. Employing a statistical model to learn these patterns by classifying the class based on the outputs of the previous models would thus allow said ML model to learn these nuanced differences between the transformers.

3.2 Experimental hardware and software

The experiments were executed on an NVidia Tesla K80 GPU which has 4992 CUDA cores and 24 GB of GDDR5 memory via the Google Colab platform. The Transformers were implemented via the KTrain library (Maiya 2020), which is a back-end for TensorFlow (Abadi et al. 2015) Keras (Chollet et al. 2015). The pretrained weights for the Transformers prior to fine-tuning were from the HuggingFace NLP Library (Wolf et al. 2019). The pre-trained T5 paraphrasing model weights were from (Chang 2020). The model was implemeted with the HuggingFace NLP Library (Wolf et al. 2019) via PyTorch (Paszke et al. 2019) and was trained for two epochs (~20 h) on the p2.xlarge AWS ec2.

The statistical models for the stacked generalisation ensemble results were implemented in Python via the Scikit-learn toolkit (Pedregosa et al. 2011) and executed on an Intel Core i7 Processor (3.7 GHz).

4 Results

Table 3 shows the overall results for all of the experiments. Every single model, even the weakest XLM for this particular problem, was improved when training on the human data alongside the augmented data which can be seen for the increases in metrics in Table 4. This required a longer training time due to the more computationally intense nature of training on a larger dataset. T5 paraphrasing for data augmentation led to an average accuracy increase of 4.01 points, and the precision, recall, and F1 scores were also improved at an average of 0.05, 0.05, and 0.07, respectively.

Interestingly, although the results strongly suggest that paraphrased data augmentation improves training, the readability of the paraphrased data was relatively mixed and some strange occurrences took place. For example, "Can

Table 4 Observed increases in training metrics for each model due to data augmentation via paraphrasing the training dataset

Model	Increase of metrics			
	Acc.	*Prec.*	*Rec.*	*F1*
BERT	8.3	0.06	0.09	0.09
DistilBERT	1.04	0.01	0.01	0.01
DistilRoBERTa	3.11	0.03	0.04	0.04
RoBERTa	1.03	0.01	0.01	0.01
XLM	1.12	0.13	0.01	0.12
XLM-RoBERTa	10.79	0.09	0.11	0.11
XLNet	2.69	0.03	0.11	0.12
Average	4.01	0.05	0.05	0.07

you stay a while and talk with me?" and "Would you mind to speak with me for a little bit? Or would that be a problem?" are perfectly reasonable requests for a conversation. But, some data such as "I want to talk to you. I am a university student. I'd just like to speak with you. I have everything to give!" is obviously an unnatural utterance, and yet also evidently contains some useful knowledge for the model to learn. Likewise, this can be noted for other classes. To give another example, "If you know British Sign Language then I would prefer to use it." was produced by the paraphrasing model, and this indeed makes sense and is a useful utterance. Similarly to the previous example, there were strange suggestions by the model such as "I want to sign but don't want to speak. Do you know the signs of a sign?" and "Why do we speak in leap motion without any real thought?". Though these sentences contain useful knowledge as can be seen from the increase in classification metrics, this suggests future work may be required to clean the augmented data (reducing the dataset by culling a selection of the worst outputs) which may lead to better performance. This would also lead to a less computationally expensive approach given that there would be fewer training examples with only those in utmost quality retained. These occurrences also suggest that although paraphrasing is useful for data augmentation when training to understand human utterances, it would be logical to not use such a model for data that is going to be presented to the user such as the chatbot's responses, given that not all paraphrased data makes sense from an English language perspective. Additionally, although it did not occur in the paraphrasing of this dataset, questions on Quora (which the T5 is trained on) can be of a sexual nature and as such thus may lead to inappropriate utterances by the chatbot.

The best performing model was RoBERTa when training on the human training set as well as the augmented data. This model achieved 98.96% accuracy with 0.99 precision, recall and F1 score. The alternative to training only on the human data achieved 97.93% accuracy with stable precision, recall and F1 scores of 0.98. The second best performing models were both the distilled version of RoBERTa and BERT, which achieved 98.55% and likewise 0.99 for the other three metrics. Interestingly, some models saw a drastic increase in classification ability when data augmentation was implemented; the BERT model rose from 90.25% classification accuracy with 0.93 precision, 0.9 recall and 0.9 F1 score with a +8.3% increase and then more stable metrics of 0.99 each as described previously. In the remainder of this section, the 98.96% performing RoBERTa model when trained upon human and T5 data is explored further. This includes, exploration of errors made overall and per specific examples, as well as an exploration of top features within successful predictions made.

Figure 7 shows a comparison between the model performance and number of trainable parameters. Note that the most complex model scored the least in terms of classification ability. The best performing model was the second most complex model of all. The least complex model, DistilBERT, achieved a relatively high accuracy of 98.34%.

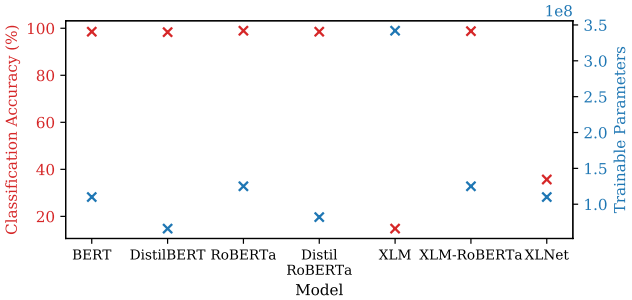

Fig. 7 Comparison of each model's classification ability and number of million trainable parameters within them

Table 5 Per-class precision, recall, and F1 score metrics for the best model

Class	Prec.	Rec.	F1
CHAT	1.00	0.99	0.99
EEG-EMOTIONS	0.99	0.97	0.98
EEG-MENTAL-STATE	0.99	1.00	0.99
JOKE	0.98	0.98	0.98
SCENE-CLASSIFICATION	1.00	1.00	1.00
SENTIMENT-ANALYSIS	0.97	0.99	0.98
SIGN-LANGUAGE	1.00	1.00	1.00

4.1 Exploration of the best transformer model

In this section, we explore the best model. The best model, as previously discussed, was the RoBERTa model when training on both the collected training data and the paraphrased data generated by the T5 model.

Table 5 shows the classification metrics for each individual class by the RoBERTa model. The error matrix for the validation data can be seen in Fig. 8. The tasks of EEG mental state classification, scene recognition, and sign language were classified perfectly. Of the imperfect classes, the task of conversational AI ('CHAT') was sometimes misclassified as a request for a joke, which is likely due to the social nature of the two activities. EEG emotional state classification was rarely mistakenly classified as the mental state recognition and sentiment analysis tasks, firstly due to the closely related EEG tasks and secondly as sentiment analysis since data often involved terms synonymous with valence or emotion. Similarly, the joke class was also rarely misclassified as sentiment analysis, for example, "tell me something funny" and "can you read this email and tell me

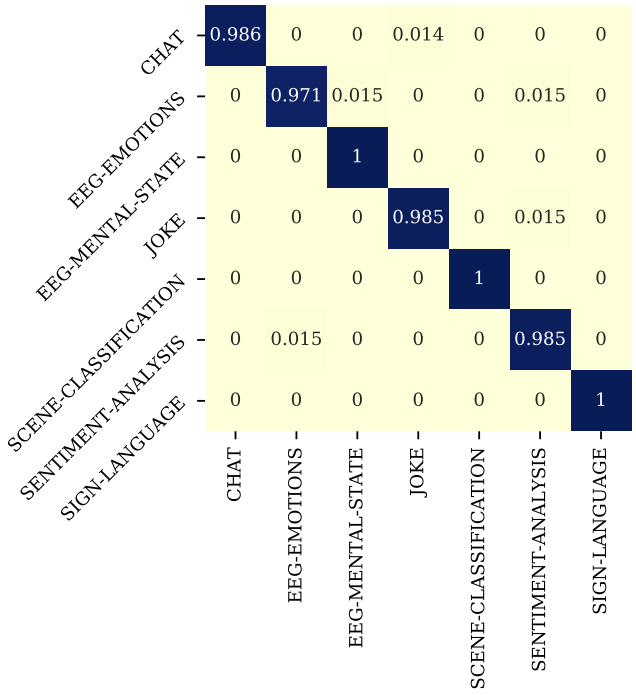

Fig. 8 Normalised confusion matrix for the best command classification model, which was RoBERTa when trained on human data and augmented T5 paraphrased data

if they are being funny with me?" ('funny' in the second context being a British slang term for sarcasm). The final class with misclassified instances was sentiment analysis, as emotional state recognition, for the same reason previously described when the error occurred vice-versa.

4.2 Mistakes and probabilities

In this section, we explore the biggest errors made when classifying the validation set by considering their losses.

Table 6 shows the most confusing data objects within the training set and Fig. 9 explores which parts of the phrase the model focused on to derive these erroneous classifications. Overall, only five misclassified sentences had a loss above 1; the worst losses were in the range of 1.05 to 6.24. The first phrase, "what is your favourite one liner?", may likely have caused confusion due to the term "one liner" which was not present within the training set. Likewise, the term "valence" in "What is the valence of my brainwaves?" was also not present within the training set, and the term "brainwaves" was most common when referring to mental state recognition rather than emotional state recognition. An interesting error occurred from the command "Run emotion classification", where the classification was incorrectly given as EEG emotional state recognition rather than Sentiment Analysis. The command collected from a human subject was ambiguous, and as such the two most likely classes were the incorrect EEG Emotions at a probability of 0.672 and the correct Sentiment Analysis at a probability of 0.32. This raises an issue to be explored in future works, given the nature of natural social interaction, it is likely that ambiguity will be present during conversation. Within this erroneous classification, two classes were far more likely than all other classes present, and thus a choice between the two in the form of a question akin to human deduction of ambiguous language would likely solve such problems and increase accuracy. Additionally, this would rarely incur the requirement of further effort from the user.

4.3 Top features within unseen data

Following the training of the model, this section explores features within data when an unseen phrase or command is uttered. That is, the examples given in this section were not data within the training or validation datasets, and thus are more accurate simulations of the model within a real-world scenario given new data to process based on the rules learnt during training.

In this regard, Fig. 10 shows an example of a correct prediction of unseen data class, for each class. Interestingly, the model shows behaviour reminiscent of human reading (Biedert et al. 2012; Kunze et al. 2013) due to transformers not being limited to considering a temporal sequence in chronological order of appearance. In the first example the most useful features were 'time to speak' followed by 'got', 'to' and 'me'. The least useful features were 'right now', which alone would be classified as 'SCENE-CLASSIFICATION' with a probability of 0.781 due to many provided training examples for such class containing questions such as 'where are you **right now**? Can you run scene recognition and tell me?'. The second example also had a strong negative impact from the word 'read' which alone would be classified as 'SENTIMENT-ANALYSIS' with a probability of 0.991 due to the existence of phrases such as 'please **read** this message **and tell me** if they are angry with me' being popular within the gathered human responses and as such the augmented data. This example found correct classification due to the terms 'emotions' and 'mind' primarily, followed by 'feeling'. Following these two first examples, the remaining five examples were strongly classified. In the mental state recognition task, even though the term 'mental state' was specifically uttered, the term 'concentrating' was the strongest feature within the statement given the goal of the algorithm to classify concentrating and relaxed states of mind. As could be expected, the 'JOKE' task was best classified by the term 'joke' itself being present, but, interestingly, the confidence of classification was increased with the phrases 'Feeling sad today.' and 'cheer me up'. The scene classification task was confidently predicted with a probability of 1 mainly due to the terms 'look around' and

Table 6 The most confusing sentences according to the model (all of those with a loss > 1) and the probabilities as to which class they were predicted to belong to

Text	"What is your favourite one liner?"						
Actual	C4						
Predicted	C6						
Loss	6.24						
Prediction probabilities	***C1***	***C2***	***C3***	***C4***	***C5***	***C6***	***C7***
	0.0163	0.001	0	0.002	0.001	0.977	0.002
Text	"What is your favourite movie?"						
Actual	C1						
Predicted	C4						
Loss	2.75						
Prediction probabilities	***C1***	***C2***	***C3***	***C4***	***C5***	***C6***	***C7***
	0.064	0.0368	0.007	0.513	0.338	0.022	0.02
Text	"How do I feel right now?"						
Actual	C1						
Predicted	C4						
Loss	2.75						
Prediction probabilities	***C1***	***C2***	***C3***	***C4***	***C5***	***C6***	***C7***
	0.007	0.01	0.352	0.434	0.016	0.176	0.005
Text	"Run emotion classification"						
Actual	C6						
Predicted	C2						
Loss	1.71						
Prediction probabilities	***C1***	***C2***	***C3***	***C4***	***C5***	***C6***	***C7***
	0	0.672	0.001	0.002	0.004	0.32	0
Text	"What is the valence of my brainwaves?"						
Actual	C2						
Predicted	C3						
Loss	1.05						
Prediction probabilities	***C1***	***C2***	***C3***	***C4***	***C5***	***C6***	***C7***
	0.001	0.349	0.647	0.001	0.001	0.002	0

Key—C1: CHAT, C2: EEG-EMOTIONS, C3: EEG-MENTAL-STATE, C4: JOKE, C5: SCENE-RECOGNITION, C6: SENTIMENT-ANALYSIS, C7: SIGN-LANGUAGE

'where you are'. The red highlight for the word 'if' alone would be classified as 'SENTIMENT-ANALYSIS' with a probability of 0.518 given the popularity of phrases along the lines of '**if** they are *emotion* or *emotion*'. The sentiment analysis task was then, again, confidently classified correctly with a probability of 1. This was due to the terms 'received this email', 'if', and 'sarcastic' being present. Finally, the sign language task was also classified with a probability of 1 most due to the features 'voice' and 'sign'. The red features highlighted, 'speaking with please' would alone be classified as 'CHAT' with a probability of 0.956, since they are strongly reminiscent to commands such as, 'can we speak about something please?'. An interesting behaviour to note from these examples is the previously described nature of reading. Transformer models are advancing the field of NLP in part thanks due to their lack of temporal restriction, ergo the limitations existent within models such as Recurrent or Long Short Term Memory Neural Networks. This allows for behaviours more similar to a human being, such as when someone may focus on certain key words first before glancing backwards for more context. Such behaviours are not possible with sequence-based text classification techniques.

4.4 Transformer ensemble results

Following the previous findings, the five strongest models which were BERT (98.55%), DistilBERT (98.34%), RoBERTa (98.96%), Distil-RoBERTa (98.55%), and

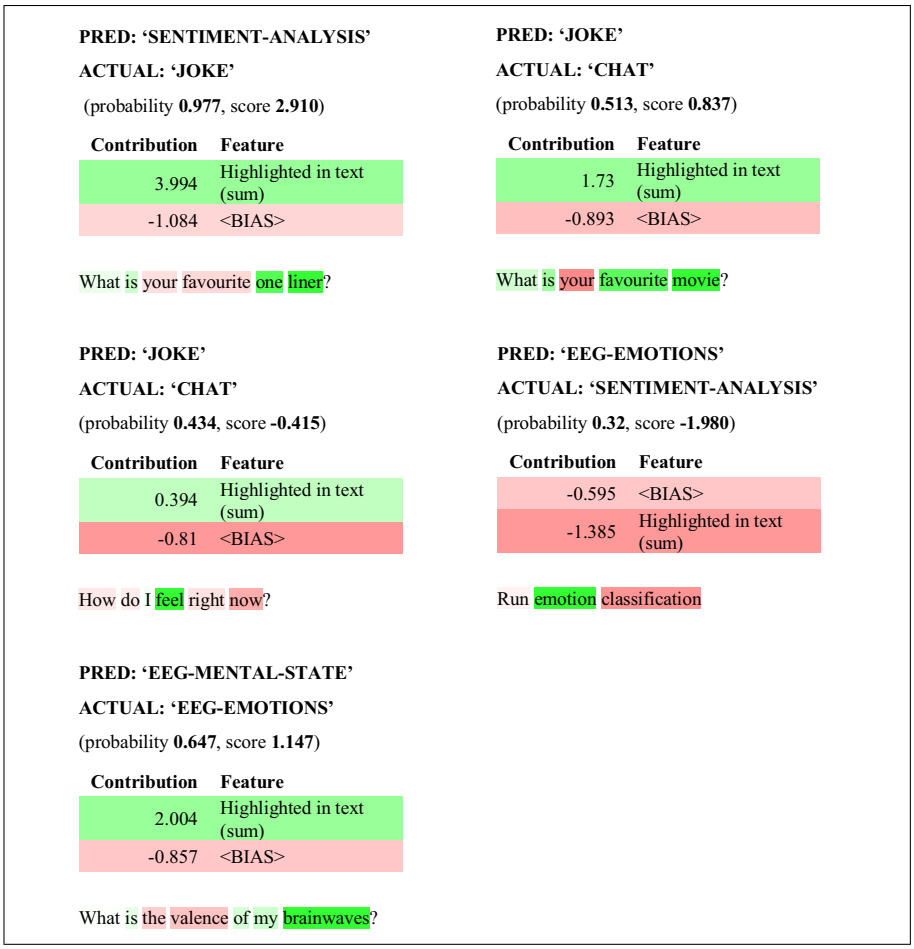

Fig. 9 Exploration and explanation for the errors made during validation which had a loss > 1 (five such cases)

XLM-RoBERTa (98.76%) are combined into a preliminary ensemble strategy as previously described. XLM (14.81%) and XLNet (35.68%) are omitted due to their low classification abilities. As noted, it was observed previously that the best score by a single model was RoBERTa which scored 98.96% classification accuracy, and thus the main goal of the statistical ensemble classifier is to learn patterns that could possibly account for making up some of the 1.04% of errors and correct for them. Initially, Table 7 shows the information gain rankings of each predictor by 10 fold cross validation on the training set alone, interestingly BERT is ranked the highest with an information gain of 2.717 (± 0.002). Following this, the results in Table 8 show the results for multiple statistical methods of ensembling the predictions of the five Transformer models (with the best performing approaches highlighted in bold); all of the models with the exception of Gaussian Naïve Bayes could outperform the best single Transformer model by an accuracy increase of at least 0.42 points. The two best models which achieved the same score were Logistic Regression and Random Forests, which when ensembling the predictions of the five transformers, could increase the accuracy by 0.63 points over RoBERTa and achieve an accuracy of 99.59%.

Finally, Fig. 11 shows the confusion matrix for both the Logistic Regression and Random Forest methods of ensembling Transformer predictions since the errors made by both models were identical. Many of the errors have been mitigated through ensembling the transformer models, with minor confusion occuring between the 'CHAT' and 'JOKE' classes and the 'SENTIMENT ANALYSIS' and 'EEG-EMOTIONS' classes.

5 Conclusion and future work

The studies performed in this work have shown primarily that data augmentation through transformer-based paraphrasing via the T5 model have positively useful effects on many state-of-the-art language transformer-based classification models. BERT and DistilBERT, RoBERTa and DisilRoBERTa, XLM, XLM-RoBERTa, and XLNet all showed increases in learning performance when learning with augmented data from the training set when compared to learning only on the original data pre-augmentation. The best single model found

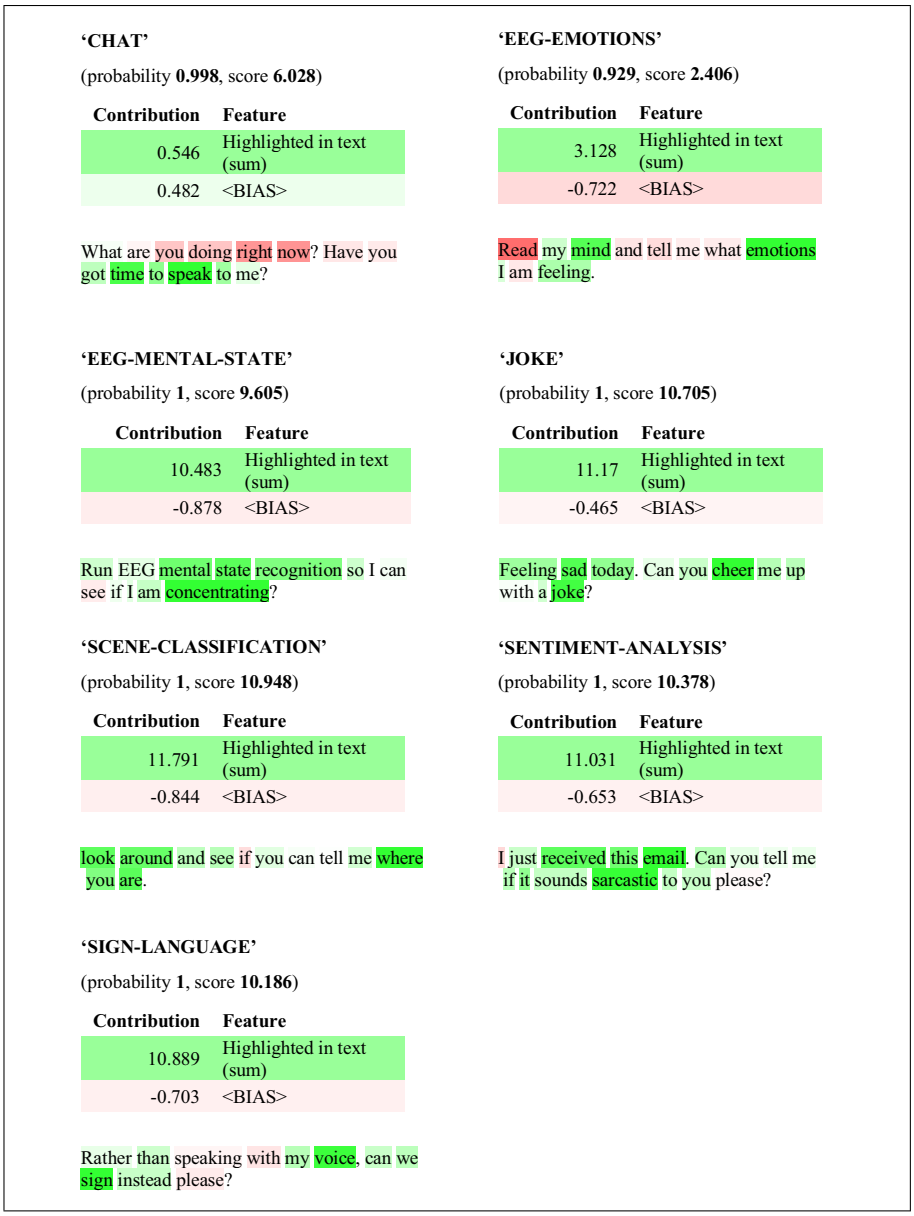

Fig. 10 Exploration of the best performing model by presenting unseen sentences and explaining predictions. Green denotes useful features and red denotes features useful for another class (detrimental to probability)

Table 7 Information Gain ranking of each predictor model by 10 fold cross validation on the training set

Predictor model (transformer)	Average ranking	Information Gain of predictions
BERT	1 (± 0)	2.717 (± 0.002)
DistilBERT	2 (± 0)	2.707 (± 0.002)
DistilRoBERTa	3.1 (± 0.3)	2.681 (± 0.001)
RoBERTa	3.9 (± 0.3)	2.676 (± 0.003)
XLM-RoBERTa	5 (± 0)	2.653 (± 0.002)

was RoBERTa, which could classify human commands to an artificially intelligent system at a rate of 98.96% accuracy, where errors were often due to ambiguity within human language. A statistical ensemble of the five best transformer models then led to an increase accuracy of 99.59% when using either Logistic Regression or a Random Forest to process the output predictions of each transformer, utilising small differences between the models when trained on the dataset. Given that several related works present XLM as a strong candidate for different language-based problems with a focus on multilingual training, it is possibly the case that there is not enough data to fine-tune XLM away from consideration of multiple languages and this leads to weak results when working with only English language. Thus in future when several languages may be considered as input to the system, XLM could be revisited in order to explore this conjecture. Although XLM did not perform well, the promising performance of XLM-RoBERTa showed that models trained on a task do not necessarily under

Table 8 Results for the ensemble learning of Transformer predictions compared to the best single model (RoBERTa)

Ensemble method	Accuracy	Precision	Recall	F1	Difference over RoBERTa
Logistic Regression	**99.59**	**0.996**	**0.996**	**0.996**	**0.63**
Random Forest	**99.59**	**0.996**	**0.996**	**0.996**	**0.63**
Multinomial Naïve Bayes	99.38	0.994	0.994	0.994	0.42
Bernoulli Naïve Bayes	99.38	0.994	0.994	0.994	0.42
Linear Discriminant Analysis	99.38	0.994	0.994	0.994	0.42
XGBoost	99.38	0.994	0.994	0.994	0.42
Support Vector Classifier	99.38	0.994	0.994	0.994	0.42
Bayesian Network	99.38	0.994	0.994	0.994	0.42
Gaussian Naïve Bayes	98.55	0.986	0.985	0.986	– 0.41

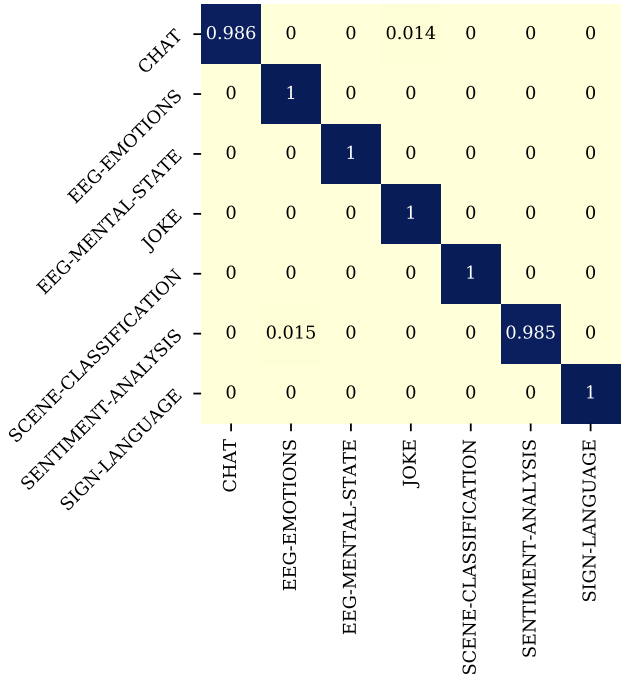

Fig. 11 Normalised confusion matrix for the best ensemble methods of Logistic Regression and Random Forest (errors made by the two were identical)

perform on another different task given the general ability of lingual understanding. With this in mind, and given that the models are too complex to train simultaneously, it may be useful in the future to explore other methods of ensembling the predictions such as the addition of the original text alongside prediction vectors, which may allow for deeper understanding behind why errors are made and allow for further NLP-based rules to overcome them. A preliminary ensemble of the five strongest models showed that classification accuracy could be further increased by treating the outputs of each transformer model as attributes in themselves, for rules to be learnt from. The experiment was limited in that attribute selection was based solely on removing the two under performing models; in future, exploration could be performed into attribute selection to fine-tune the number of models used as input. Additionally, only a predicted labels in the form of nominal attributes were used as input, whereas additional attributes such as probabilities of each output class could be utilised in order to provide more information for the statistical ensemble classifier. The data in this work was split 70/30 and paraphrasing was executed on the 70% of training data only in order not to expose a classification model to paraphrased text of data contained in the testing set. This is performed in order to prevent training data possibly baring strong similarity to test data (since the output of the T5 may or may not be very similar to the input, and is difficult to control in this regard). In future, metrics such as the accuracy, precision, recall, and F1 scores etc. could be made more scientifically accurate based on the knowledge gained from this study by performing K-fold Cross Validation or even Leave One Out Cross Validation if the computational resources are available to do so.

6 Ethics

All users who answered the questionnaire agreed to the following statement:

The data collected from this form will remain completely anonymous and used for training a transformation-based chatbot. The more examples of a command or statement the bot can observe, the more accurate it will be at giving the correct response. The responses will be expanded by exploring paraphrases of answers and then further transformed by a model pre-trained on a large corpus of text and fine-tuned on the goal-based statements and requests given here.

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X (2015) TensorFlow: large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/, software available from tensorflow.org

Bannard C, Callison-Burch C (2005) Paraphrasing with bilingual parallel corpora. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pp 597–604

Biedert R, Hees J, Dengel A, Buscher G (2012) A robust realtime reading-skimming classifier. In: Proceedings of the Symposium on Eye Tracking Research and Applications, pp 123–130

Bird JJ, Ekárt A, Faria DR (2018a) Learning from interaction: An intelligent networked-based human-bot and bot-bot chatbot system. In: UK workshop on computational intelligence, Springer, pp 179–190

Bird JJ, Manso LJ, Ribeiro EP, Ekart A, Faria DR (2018b) A study on mental state classification using EEG-based brain-machine interface. In: 2018 International conference on intelligent systems (IS), IEEE, pp 795–800

Bird JJ, Ekart A, Buckingham C, Faria DR (2019a) Mental emotional sentiment classification with an EEG-based brain-machine interface. In: Proceedings of the International Conference on Digital Image and Signal Processing (DISP'19)

Bird JJ, Ekart A, Buckingham CD, Faria DR (2019b) High resolution sentiment analysis by ensemble classification. In: Intelligent computing-proceedings of the computing conference, Springer, pp 593–606

Bird JJ, Ekárt A, Faria DR (2020a) British sign language recognition via late fusion of computer vision and leap motion with transfer learning to American sign language. Sensors 20(18):5151

Bird JJ, Faria DR, Premebida C, Ekárt A, Vogiatzis G (2020b) Look and listen: a multi-modality late fusion approach to scene classification for autonomous machines. arXiv preprint arXiv:200710175

Bollweg L, Kurzke M, Shahriar KA, Weber P (2018) When robots talk-improving the scalability of practical assignments in MOOCS using chatbots. In: EdMedia+ innovate learning, association for the advancement of computing in education (AACE), pp 1455–1464

Candello H, Pinhanez C, Pichiliani MC, Guerra MA, Gatti de Bayser M (2018) Having an animated coffee with a group of chatbots from the 19th century. In: Extended abstracts of the 2018 CHI conference on human factors in computing systems, pp 1–4

Chada R (2020) Simultaneous paraphrasing and translation by fine-tuning transformer models. arXiv preprint arXiv:200505570

Chang E (2020) Ellachang/T5-paraphraser: modified version of Google's t5 model that produces paraphrases of a given input sentence. https://github.com/EllaChang/T5-Paraphraser, (Accessed on 08/11/2020)

Chang WC, Yu HF, Zhong K, Yang Y, Dhillon I (2019) X-bert: extreme multi-label text classification with using bidirectional encoder representations from transformers. arXiv preprint arXiv:190502331

Chollet F, et al. (2015) Keras. https://keras.io

Ciechanowski L, Przegalinska A, Magnuski M, Gloor P (2019) In the shades of the uncanny valley: an experimental study of human-chatbot interaction. Futur Gener Comput Syst 92:539–548

Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V (2019) Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:191102116

Conneau A, Lample G (2019) Cross-lingual language model pretraining. In: Advances in Neural Information Processing Systems, pp 7059–7069

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805

Devlin J, Chang MW (2018) Open sourcing bert: State-of-the-art pre-training for natural language processing. Google AI Blog Weblog. https://aigoogleblog.com/2018/11/open-sourcing-bertstate-of-art-pre html (Accessed 4 Dec 2019)

Di Gangi MA, Negri M, Turchi M (2019) Adapting transformer to end-to-end spoken language translation. In: INTERSPEECH 2019, International Speech Communication Association (ISCA), pp 1133–1137

Dimovski M, Musat C, Ilievski V, Hossmann A, Baeriswyl M (2018) Submodularity-inspired data selection for goal-oriented chatbot training based on sentence embeddings. arXiv preprint arXiv:180200757

Eckstein G, Schramm W, Noxon M, Snyder J (2019) Reading l1 and l2 writing: an eye-tracking study of tesol rater behavior. TESL-EJ 23(1):n1

Goodfellow I, Bengio Y, Courville A (2017) Deep learning (adaptive computation and machine learning series). Cambridge Massachusetts, pp 321–359

Haller E, Rebedea T (2013) Designing a chat-bot that simulates an historical figure. In: 2013 19th international conference on control systems and computer science, IEEE, pp 582–589

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Hou Y, Liu Y, Che W, Liu T (2018) Sequence-to-sequence data augmentation for dialogue language understanding. arXiv preprint arXiv:180701554

Howard J, Ruder S (2018) Universal language model fine-tuning for text classification. arXiv preprint arXiv:180106146

Jin L, King D, Hussein A, White M, Danforth D (2018) Using paraphrasing and memory-augmented models to combat data sparsity in question interpretation with a virtual patient dialogue system. In: Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications, pp 13–23

Kerlyl A, Hall P, Bull S (2006) Bringing chatbots into education: towards natural language negotiation of open learner models. In: International Conference on Innovative Techniques and Applications of Artificial Intelligence, Springer, pp 179–192

Kruger JL, Steyn F (2014) Subtitles and eye tracking: reading and performance. Read Res Q 49(1):105–120

Kunze K, Ishimaru S, Utsumi Y, Kise K (2013) My reading life: towards utilizing eyetracking on unmodified tablets and phones. In: Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication, pp 283–286

Leonhardt MD, Tarouco L, Vicari RM, Santos ER, da Silva MdS (2007) Using chatbots for network management training through problem-based oriented education. In: Seventh IEEE international conference on advanced learning technologies (ICALT 2007), IEEE, pp 845–847

Lewis M, Ghazvininejad M, Ghosh G, Aghajanyan A, Wang S, Zettlemoyer L (2020) Pre-training via paraphrasing. arXiv preprint arXiv:200615020

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:190711692

Lukovnikov D, Fischer A, Lehmann J (2019) Pretrained transformers for simple question answering over knowledge graphs. In: International semantic web conference, Springer, pp 470–486

Maiya AS (2020) ktrain: a low-code library for augmented machine learning arXiv arXiv:2004.10703 [cs.LG]

Manurung R, Ritchie G, Pain H, Waller A, O'Mara D, Black R (2008) The construction of a pun generator for language skills development. Appl Artif Intell 22(9):841–869

Marquis A, Kaabi MA, Leung T, Boush F (2020) What the eyes hear: an eye-tracking study on phonological awareness in emirati arabic. Front Psychol 11:452

Marton Y, Callison-Burch C, Resnik P (2009) Improved statistical machine translation using monolingually-derived paraphrases. In: Proceedings of the 2009 conference on empirical methods in natural language processing, pp 381–390

Naseem U, Razzak I, Musial K, Imran M (2020) Transformer based deep intelligent contextual embedding for twitter sentiment analysis. Future Gener Comput Syst 113:58–69. https://doi.org/10.1016/j.future.2020.06.050

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) Pytorch: An imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems 32, Curran Associates, Inc., pp 8024–8035, http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

Petrović S, Matthews D (2013) Unsupervised joke generation from big data. In: Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: short papers), pp 228–232

Qi D, Su L, Song J, Cui E, Bharti T, Sacheti A (2020) Imagebert: cross-modal pre-training with large-scale weak-supervised image-text data. arXiv preprint arXiv:200107966

Quora (2017) Quora question pairs | kaggle. https://www.kaggle.com/c/quora-question-pairs, (Accessed on 08/11/2020)

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. OpenAI blog 1(8):9

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2019) Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:191010683

Roller S, Boureau YL, Weston J, Bordes A, Dinan E, Fan A, Gunning D, Ju D, Li M, Poff S et al (2020) Open-domain conversational agents: current progress open problems, and future directions. arXiv preprint arXiv:200612442

Sanh V, Debut L, Chaumond J, Wolf T (2019) Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:191001108

Schmidhuber J (1992) Learning complex, extended sequences using the principle of history compression. Neural Comput 4(2):234–242

Shagass C, Roemer RA, Amadeo M (1976) Eye-tracking performance and engagement of attention. Arch Gen Psychiatry 33(1):121–125

Shangipour ataei T, Javdan S, Minaei-Bidgoli B (2020) Applying transformers and aspect-based sentiment analysis approaches on sarcasm detection. In: Proceedings of the second workshop on figurative language processing, association for computational linguistics, pp 67–71, https://doi.org/10.18653/v1/2020.figlang-1.9, https://www.aclweb.org/anthology/2020.figlang-1.9

Shao T, Guo Y, Chen H, Hao Z (2019) Transformer-based neural network for answer selection in question answering. IEEE Access 7:26146–26156

Shleifer S (2019) Low resource text classification with ulmfit and backtranslation. arXiv preprint arXiv:190309244

Stephens KR (2002) What has the loebner contest told us about conversant systems. Retrieved August 2021

Sun C, Qiu X, Xu Y, Huang X (2019) How to fine-tune bert for text classification? In: China National Conference on Chinese Computational Linguistics, Springer, pp 194–206

Sun L, Xia C, Yin W, Liang T, Yu PS, He L (2020) Mixup-transfomer: Dynamic data augmentation for nlp tasks. arXiv preprint arXiv:201002394

Tenney I, Das D, Pavlick E (2019) Bert rediscovers the classical nlp pipeline. arXiv preprint arXiv:190505950

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008

Virkar M, Honmane V, Rao SU (2019) Humanizing the chatbot with semantics based natural language generation. In: 2019 International Conference on Intelligent Computing and Control Systems (ICCS), IEEE, pp 891–894

Wang CC, Hung JC, Chen SN, Chang HP (2019) Tracking students' visual attention on manga-based interactive e-book while reading: an eye-movement approach. Multimedia Tools Appl 78(4):4813–4834

Wang Q, Li B, Xiao T, Zhu J, Li C, Wong DF, Chao LS (2019b) Learning deep transformer models for machine translation. arXiv preprint arXiv:190601787

Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR (2018) Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:180407461

Wei J, Zou K (2019) Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:190111196

Williams RJ, Zipser D (1989) A learning algorithm for continually running fully recurrent neural networks. Neural Comput 1(2):270–280

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Brew J (2019) Huggingface's transformers: state-of-the-art natural language processing. arXiv:1910.03771

Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV (2019) Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in neural information processing systems, 32

Zhang C, Ma Y (2012) Ensemble machine learning: methods and applications. Springer, Berlin

Zhang J, Luan H, Sun M, Zhai F, Xu J, Zhang M, Liu Y (2018) Improving the transformer translation model with document-level context. arXiv preprint arXiv:181003581