2022

Artificial Intelligence

Natural Language Processing

Next Argument Prediction with Transformers

SAMI MOHAMMED OSMAN

BOLOGNA UNIVERSITY | Italy

Contents

1 Summary

Natural Language Understanding is an ongoing research field in NLP that is trying to achieve superhuman performance on tasks such as Next Sentence Prediction, text classification, machine translation and automated reasoning. The real application is quite important: fake-news checking, assistant-application for people with disabilities, assistant for law documentation such as term of service and so on. In NLP, starting from non-neural models, performance was increased using neural model such as linear, convolution and recurrent networks (RNNs, LSTMs, CNNs etc.) reaching its peak with attention mechanisms.

Next sentence prediction is one of the many tasks in Natural language understanding which can be further seen as tasks like:

1. The goal to identify whether the second sentence is entailment, contradiction, or neutral with respect to the first sentence.
2. The goal to determine whether two questions are semantically equal.
3. The task to determine whether the second sentence is the continuation of the first or not.

In my work, I used a Transformer-based architecture to solve a task that determine whether the second argument is the entailment of the first or not, using a dataset from Kialo.com website. More specifically, I fine-tuned a pre-trained BERT-based model to extract an embedding from the two argument pairs to find the probability of argument1 precedes argument 2. I did some experiments to try to improve the performances of the baseline by extending its architecture with LSTM/GRU encoder-decoder and by changing the pre-trained Bert model as embeddings. The models were then compared using F1 and Accuracy scores computed on the same test set.

2. Background.

2.1 Transformers:

"Attention is all you need"

The "Attention is all you need" paper has introduced a game-changing architecture: the Transformer, which exploits a self-attention mechanism to solve NLP tasks without the need of using recurrent models.

Until the advent of Transformers, RNN-based architectures allowed to achieve pretty good results on a variety of NLP tasks. RNN encoder-decoder architectures with attention mechanisms are still performing very well if accurately designed and trained. The problem with RNNs, however, is in their sequential nature, which does not allow to exploit the parallelism offered by GPUs, inhibiting parallelization. Due to this problem, big RNNs are slow to train. long range information is lost because of the bottleneck problem.

The parallelization problem was partly solved with CNNs and with convolutional sequence models. Furthermore, to avoid the bottleneck problem, attention mechanisms and residual connections were implemented. On the other hand, Transformers do not have recurrent layers and for this reason they can perfectly exploit GPU parallelization, allowing faster training time, and because they don't use a sequential approach like the RNNs, they can perfectly capture long range dependencies without the bottleneck problem. Thanks to parallelization, the Transformer can be trained in a fraction of the time with respect to recurrent networks (which require the state of the model at previous timestep), reaching state-of-the-art performance in a more efficient way.

The Transformer is an encoder-decoder where Multi-Head Self Attention layers allow the network to focus on specific positions in the input. The encoder and decoder are repeated, and each decoder receive as input the output of the last encoder. What makes this conceptually so much more appealing than some LSTM cell is that we can physically see a separation in tasks: for example, in an English to French translation task, the encoder learns what is English, what grammar is and more importantly what the context is, while the decoder learns how English words are related to French words.

Noticeable layers in the Transformer architecture are:

- ✓ *Multi-Head Self Attention layer*, which contains multiple Self Attention mechanisms, where a weighted similarity measure (in general dot product) is computed on the input and then a SoftMax computes probabilities, summing up to 1, that should capture attention information on different parts of the input
- ✓ *Positional Encoding layer,* which encodes each i-th embedding of each p-th token position.

The Transformer architecture is used in many NLP tasks and researchers achieved state-of-the-art results in those tasks in reasonable amount of training time. Moreover, the pretrained models from google and Hugging Face like Bert-base, DistilBert, albert etc. made it more accessible and affordable to get the language models of the Natural Language Processing.

2.2 BERT:

Researchers also came up with novel architectures like BERT, GPT, GPT-2 etc. which allowed to achieve the actual state-of-the-art results for various NLP tasks. BERT introduced by Google in 2018, uses the Transformer to learn a language model based on context vectors, allowing to solve any kind of NLP tasks.

For this reason, the training of a BERT-based system is divided in two parts:

- ✓ **Pre-training:** it consists in investing a lot of computing power to pretrain an architecture which understands the language using Masked Language Modeling (MLM) and Next Sentence Prediction (NSP).

  In this phase, a language model is learnt by masking a portion of the input text (usually around 15%), so the model learns to predict the masked words, training a model to fill the blank spaces.

  In case of the next sentence prediction, BERT takes in two sentences and determines if this second sentence follows the first in kind of what is like binary classification problem.

  This helps BERT understand the context across different sentences themselves. Using both these methods together, BERT gets a good understanding of the language it is being trained on.
- ✓ **Fine-tuning:** it consists in tuning a previously learnt architecture (or language model) on a specific target task.

  We can now "adapt" the pre-trained BERT language model on very specific NLP tasks. For example, in the Next Sentence Prediction, what we need to do is to replace the fully connected output layers of the original network with a fresh set of fully connected layers that can basically output 0 or 1 (**0** indicates “Is Next Sentence” while **1** indicates “Not Next Sentence”)

  Then, we can perform supervised training using a Paired Sentence dataset. It won’t take long since it is only the output parameters that are learned from scratch, while the rest of the model parameters are just slightly fine-tuned and as a result training time is faster.

BERT also have a special input representation: the input to BERT are two sentences A and B which are given as a unique sentence separated by a special [SEP] token, along with another special [CLS] token that marks the start of the input. It also uses a special input embedding indicating if a token in the input belongs to sentence A or to sentence B (sentence embedding), along with traditional positional embedding.

The embedding for the tokens is constructed from three vectors:

- ✓ *Token embeddings:* This is a pre-trained embedding; the main paper uses Word Piece embedding that has 30 thousand tokens.
- ✓ *Segment embeddings:* This is basically a number indicating to which sentence each token belongs.
- ✓ *Positional embeddings:* This are embeddings of the positions of the tokens in the sentences.

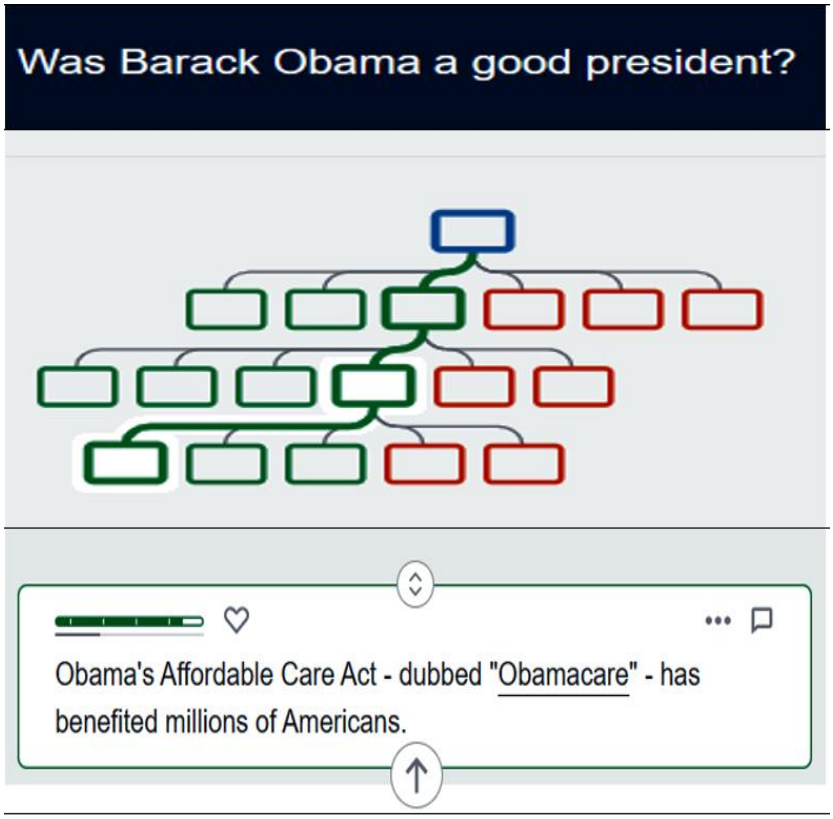
Basically, if we want to use BERT to solve the Next Sentence Prediction task, the two input sentences are, respectively, the first sentence followed by the second sentence. These inputs are used to predict whether the second sentence entails the first, so the output layer is binary classifier which predicts 0 for “Yes Next” or “Not Next”.

3 Dataset description.

3.1 Raw Data:

In my project the dataset used for training is downloaded from Kialo.com and then processed in such a way that it can be used for different NLP tasks. Kialo Edu is the world's largest argument mapping and debate site, specifically designed for classroom use. It's clear, visually compelling format makes it easy to follow the logical structure of a discussion and facilitates thoughtful collaboration.

The downloaded files have a tree like data structure where people write their argument based on the previous argument or related to the discussion topic. On top of that arguments can be cons/pros relative to the previous argument.



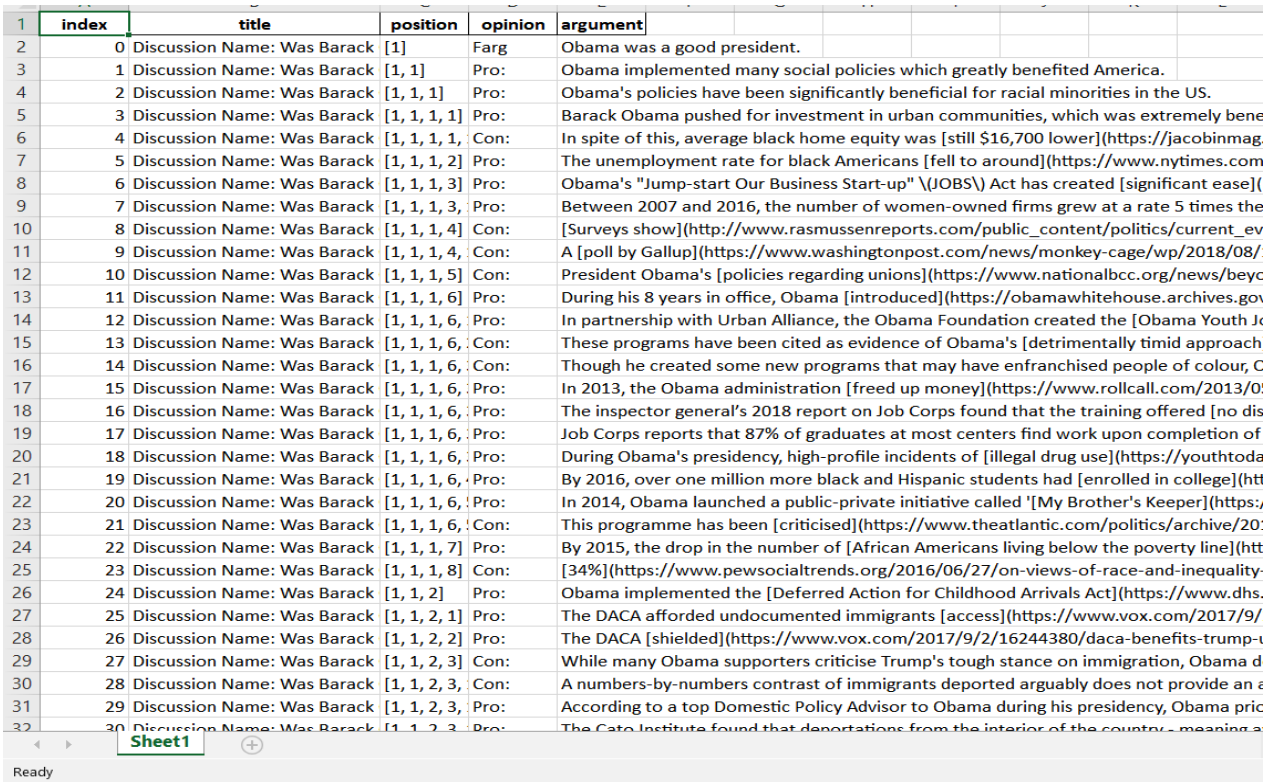

```
Discussion Name: Was Barack Obama a good president?

1. Barack Obama was a good president.
1.1. Pro: Obama implemented many social policies which greatly benefited America.
1.1.1. Pro: Obama's policies have been significantly beneficial for racial minorities in the US.
1.1.1.1. Pro: Barack Obama pushed for investment in urban communities, which was extremely benefi
1.1.1.1.1. Con: In spite of this, average black home equity was [still $16,700 lower](https://jac
1.1.1.2. Pro: The unemployment rate for black Americans [fell to around](https://www.nytimes.com/
1.1.1.3. Pro: Obama's "Jump-start Our Business Start-up" \(JOBS\) Act has created [significant ea
1.1.1.3.1. Pro: Between 2007 and 2016, the number of women-owned firms grew at a rate 5 times the
1.1.1.4. Con: [Surveys show](http://www.rasmussenreports.com/public_content/politics/current_even
1.1.1.4.1. Con: A [poll by Gallup](https://www.washingtonpost.com/news/monkey-cage/wp/2018/08/17/
1.1.1.5. Con: President Obama's [policies regarding unions](https://www.nationalbcc.org/news/beyc
1.1.1.6. Pro: During his 8 years in office, Obama [introduced](https://obamawhitehouse.archives.g
1.1.1.6.1. Pro: In partnership with Urban Alliance, the Obama Foundation created the [Obama Youth
1.1.1.6.2. Con: These programs have been cited as evidence of Obama's [detrimentally timid approa
1.1.1.6.3. Con: Though he created some new programs that may have enfranchised people of colour,
1.1.1.6.3.1. Pro: In 2013, the Obama administration [freed up money](https://www.rollcall.com/201
1.1.1.6.3.1.1. Pro: The inspector general's 2018 report on Job Corps found that the training offe
1.1.1.6.3.1.2. Pro: Job Corps reports that 87% of graduates at most centers find work upon comple
1.1.1.6.3.1.3. Pro: During Obama's presidency, high-profile incidents of [illegal drug use](https
1.1.1.6.4. Pro: By 2016, over one million more black and Hispanic students had [enrolled in colle
1.1.1.6.5. Pro: In 2014, Obama launched a public-private initiative called '[My Brother's Keeper]
1.1.1.6.5.1. Con: This programme has been [criticised](https://www.theatlantic.com/politics/archi
1.1.1.7. Pro: By 2015, the drop in the number of [African Americans living below the poverty line
```

3.2 First transformation of Data:

3.2.1 Create a general dataset:

This dataset is consisting of all posible details of the file downloaded from Kialo website and store them in pkl file with a proper Dataframe format. Takes all arguments in a file/folder and store them in .pkl file “single argument per dataset entry”.



	index	title	position	opinion	argument
2	0	Discussion Name: Was Barack	[1]	Farg	Obama was a good president.
3	1	Discussion Name: Was Barack	[1, 1]	Pro:	Obama implemented many social policies which greatly benefited America.
4	2	Discussion Name: Was Barack	[1, 1, 1]	Pro:	Obama's policies have been significantly beneficial for racial minorities in the US.
5	3	Discussion Name: Was Barack	[1, 1, 1, 1]	Pro:	Barack Obama pushed for investment in urban communities, which was extremely bene
6	4	Discussion Name: Was Barack	[1, 1, 1, 1,	Con:	In spite of this, average black home equity was [still $16,700 lower](https://jacobinmag.
7	5	Discussion Name: Was Barack	[1, 1, 1, 2]	Pro:	The unemployment rate for black Americans [fell to around](https://www.nytimes.com
8	6	Discussion Name: Was Barack	[1, 1, 1, 3]	Pro:	Obama's "Jump-start Our Business Start-up" \(JOBS\) Act has created [significant ease](
9	7	Discussion Name: Was Barack	[1, 1, 1, 3,	Pro:	Between 2007 and 2016, the number of women-owned firms grew at a rate 5 times the
10	8	Discussion Name: Was Barack	[1, 1, 1, 4]	Con:	[Surveys show](http://www.rasmussenreports.com/public_content/politics/current_ev
11	9	Discussion Name: Was Barack	[1, 1, 1, 4,	Con:	A [poll by Gallup](https://www.washingtonpost.com/news/monkey-cage/wp/2018/08/
12	10	Discussion Name: Was Barack	[1, 1, 1, 5]	Con:	President Obama's [policies regarding unions](https://www.nationalbcc.org/news/beyc
13	11	Discussion Name: Was Barack	[1, 1, 1, 6]	Pro:	During his 8 years in office, Obama [introduced](https://obamawhitehouse.archives.gov
14	12	Discussion Name: Was Barack	[1, 1, 1, 6,	Pro:	In partnership with Urban Alliance, the Obama Foundation created the [Obama Youth Jo
15	13	Discussion Name: Was Barack	[1, 1, 1, 6,	Con:	These programs have been cited as evidence of Obama's [detrimentally timid approach
16	14	Discussion Name: Was Barack	[1, 1, 1, 6,	Con:	Though he created some new programs that may have enfranchised people of colour, C
17	15	Discussion Name: Was Barack	[1, 1, 1, 6,	Pro:	In 2013, the Obama administration [freed up money](https://www.rollcall.com/2013/05
18	16	Discussion Name: Was Barack	[1, 1, 1, 6,	Pro:	The inspector general’s 2018 report on Job Corps found that the training offered [no dis
19	17	Discussion Name: Was Barack	[1, 1, 1, 6,	Pro:	Job Corps reports that 87% of graduates at most centers find work upon completion of
20	18	Discussion Name: Was Barack	[1, 1, 1, 6,	Pro:	During Obama's presidency, high-profile incidents of [illegal drug use](https://youthtoda
21	19	Discussion Name: Was Barack	[1, 1, 1, 6,	Pro:	By 2016, over one million more black and Hispanic students had [enrolled in college](htt
22	20	Discussion Name: Was Barack	[1, 1, 1, 6,	Pro:	In 2014, Obama launched a public-private initiative called '[My Brother's Keeper](https:/
23	21	Discussion Name: Was Barack	[1, 1, 1, 6,	Con:	This programme has been [criticised](https://www.theatlantic.com/politics/archive/201
24	22	Discussion Name: Was Barack	[1, 1, 1, 7]	Pro:	By 2015, the drop in the number of [African Americans living below the poverty line](htt
25	23	Discussion Name: Was Barack	[1, 1, 1, 8]	Con:	[34%](https://www.pewsocialtrends.org/2016/06/27/on-views-of-race-and-inequality-
26	24	Discussion Name: Was Barack	[1, 1, 2]	Pro:	Obama implemented the [Deferred Action for Childhood Arrivals Act](https://www.dhs.
27	25	Discussion Name: Was Barack	[1, 1, 2, 1]	Pro:	The DACA afforded undocumented immigrants [access](https://www.vox.com/2017/9/
28	26	Discussion Name: Was Barack	[1, 1, 2, 2]	Pro:	The DACA [shielded](https://www.vox.com/2017/9/2/16244380/daca-benefits-trump-u
29	27	Discussion Name: Was Barack	[1, 1, 2, 3]	Con:	While many Obama supporters criticise Trump's tough stance on immigration, Obama d
30	28	Discussion Name: Was Barack	[1, 1, 2, 3,	Con:	A numbers-by-numbers contrast of immigrants deported arguably does not provide an a
31	29	Discussion Name: Was Barack	[1, 1, 2, 3,	Pro:	According to a top Domestic Policy Advisor to Obama during his presidency, Obama pric
32	30	Discussion Name: Was Barack	[1, 1, 2, 3,	Pro:	The Cato Institute found that deportations from the interior of the country - meaning a

Sheet1

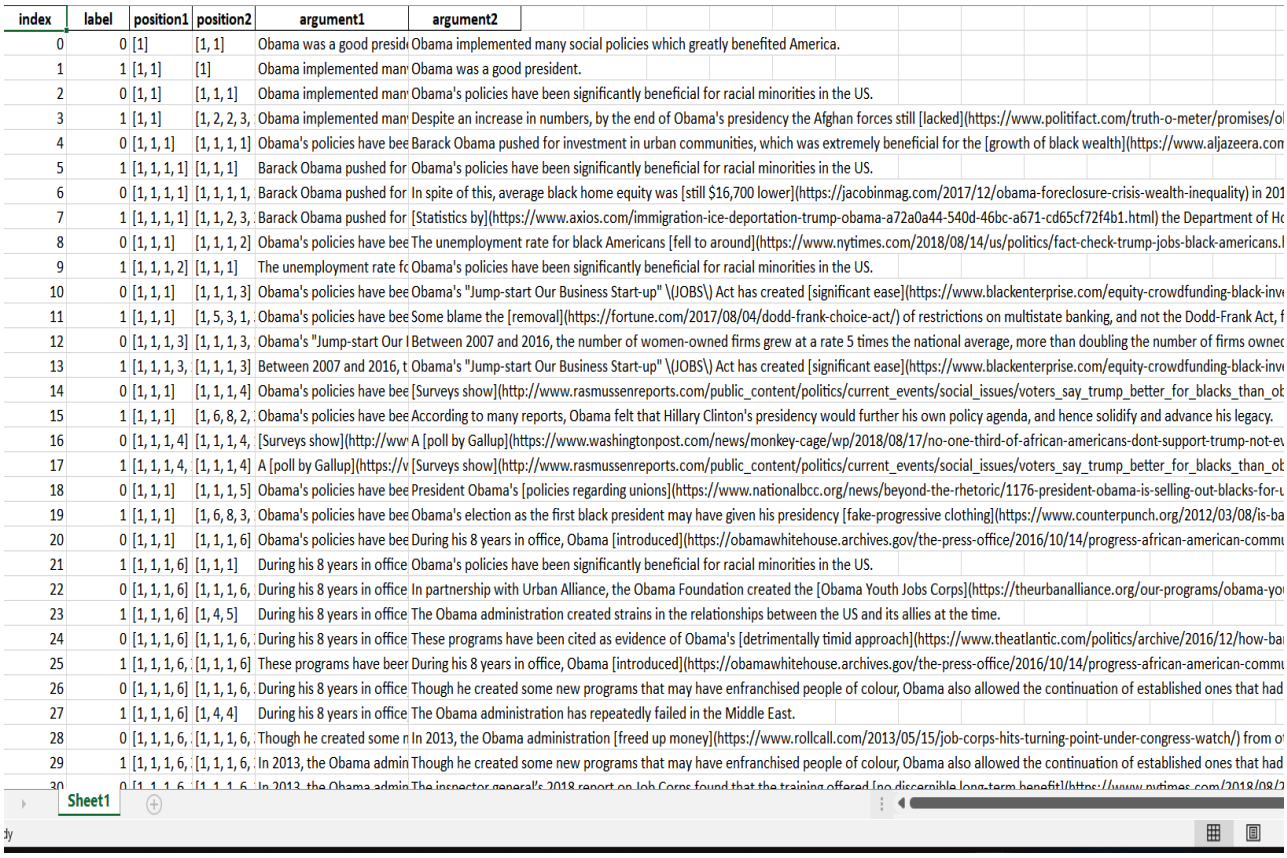
Ready

3.3 Second transformation of Data:

3.3.1 Create a dataset for Next Argument Prediction:

The dataset is consisting of a special format which is consistent to the task on hand (NSP). At this point the dataset has the structure [index, label, argument1, argument2], that I am going to use for the next argument prediction. Here, the arguments are ordered as argument 1 and argument 2 to their respective position in the file index referenced from the general dataset.

- ‘0’ label indicates argument 2 is entalilement to argument 1. 50% of the dataset.
- ‘1’ label indicates argument 2 is not entalement to argument 1. This is arangment is done using to approaches:
  - ✓ Transpose the arguments. 25% of the dataset.
  - ✓ Randomize the index of argument 2 while pairing. 25% of the dataset



index	label	position1	position2	argument1	argument2
0	0	[1]	[1, 1]	Obama was a good presid	Obama implemented many social policies which greatly benefited America.
1	1	[1, 1]	[1]	Obama implemented man	Obama was a good president.
2	0	[1, 1]	[1, 1, 1]	Obama implemented man	Obama's policies have been significantly beneficial for racial minorities in the US.
3	1	[1, 1]	[1, 2, 2, 3,	Obama implemented man	Despite an increase in numbers, by the end of Obama's presidency the Afghan forces still [lacked](https://www.politifact.com/truth-o-meter/promises/ol
4	0	[1, 1, 1]	[1, 1, 1, 1]	Obama's policies have bee	Barack Obama pushed for investment in urban communities, which was extremely beneficial for the [growth of black wealth](https://www.aljazeera.com
5	1	[1, 1, 1, 1]	[1, 1, 1]	Barack Obama pushed for	Obama's policies have been significantly beneficial for racial minorities in the US.
6	0	[1, 1, 1, 1]	[1, 1, 1, 1,	Barack Obama pushed for	In spite of this, average black home equity was [still $16,700 lower](https://jacobinmag.com/2017/12/obama-foreclosure-crisis-wealth-inequality) in 201
7	1	[1, 1, 1, 1]	[1, 1, 2, 3,	Barack Obama pushed for	[Statistics by](https://www.axios.com/immigration-ice-deportation-trump-obama-a72a0a44-540d-46bc-a671-cd65cf72f4b1.html) the Department of Ho
8	0	[1, 1, 1]	[1, 1, 1, 2]	Obama's policies have bee	The unemployment rate for black Americans [fell to around](https://www.nytimes.com/2018/08/14/us/politics/fact-check-trump-jobs-black-americans.l
9	1	[1, 1, 1, 2]	[1, 1, 1]	The unemployment rate fc	Obama's policies have been significantly beneficial for racial minorities in the US.
10	0	[1, 1, 1]	[1, 1, 1, 3]	Obama's policies have bee	Obama's "Jump-start Our Business Start-up" \(JOBS\) Act has created [significant ease](https://www.blackenterprise.com/equity-crowdfunding-black-inv
11	1	[1, 1, 1]	[1, 5, 3, 1,	Obama's policies have bee	Some blame the [removal](https://fortune.com/2017/08/04/dodd-frank-choice-act/) of restrictions on multistate banking, and not the Dodd-Frank Act, f
12	0	[1, 1, 1, 3]	[1, 1, 1, 3,	Obama's "Jump-start Our I	Between 2007 and 2016, the number of women-owned firms grew at a rate 5 times the national average, more than doubling the number of firms ownec
13	1	[1, 1, 1, 3,	[1, 1, 1, 3]	Between 2007 and 2016, t	Obama's "Jump-start Our Business Start-up" \(JOBS\) Act has created [significant ease](https://www.blackenterprise.com/equity-crowdfunding-black-inv
14	0	[1, 1, 1]	[1, 1, 1, 4]	Obama's policies have bee	[Surveys show](http://www.rasmussenreports.com/public_content/politics/current_events/social_issues/voters_say_trump_better_for_blacks_than_ob
15	1	[1, 1, 1]	[1, 6, 8, 2,	Obama's policies have bee	According to many reports, Obama felt that Hillary Clinton's presidency would further his own policy agenda, and hence solidify and advance his legacy.
16	0	[1, 1, 1, 4]	[1, 1, 1, 4,	[Surveys show](http://ww	A [poll by Gallup](https://www.washingtonpost.com/news/monkey-cage/wp/2018/08/17/no-one-third-of-african-americans-dont-support-trump-not-ev
17	1	[1, 1, 1, 4,	[1, 1, 1, 4]	A [poll by Gallup](https://v	[Surveys show](http://www.rasmussenreports.com/public_content/politics/current_events/social_issues/voters_say_trump_better_for_blacks_than_ob
18	0	[1, 1, 1]	[1, 1, 1, 5]	Obama's policies have bee	President Obama's [policies regarding unions](https://www.nationalbcc.org/news/beyond-the-rhetoric/1176-president-obama-is-selling-out-blacks-for-u
19	1	[1, 1, 1]	[1, 6, 8, 3,	Obama's policies have bee	Obama's election as the first black president may have given his presidency [fake-progressive clothing](https://www.counterpunch.org/2012/03/08/is-ba
20	0	[1, 1, 1]	[1, 1, 1, 6]	Obama's policies have bee	During his 8 years in office, Obama [introduced](https://obamawhitehouse.archives.gov/the-press-office/2016/10/14/progress-african-american-commu
21	1	[1, 1, 1, 6]	[1, 1, 1]	During his 8 years in office	Obama's policies have been significantly beneficial for racial minorities in the US.
22	0	[1, 1, 1, 6]	[1, 1, 1, 6,	During his 8 years in office	In partnership with Urban Alliance, the Obama Foundation created the [Obama Youth Jobs Corps](https://theurbanalliance.org/our-programs/obama-yo
23	1	[1, 1, 1, 6]	[1, 4, 5]	During his 8 years in office	The Obama administration created strains in the relationships between the US and its allies at the time.
24	0	[1, 1, 1, 6]	[1, 1, 1, 6,	During his 8 years in office	These programs have been cited as evidence of Obama's [detrimentally timid approach](https://www.theatlantic.com/politics/archive/2016/12/how-ba
25	1	[1, 1, 1, 6,	[1, 1, 1, 6]	These programs have beer	During his 8 years in office, Obama [introduced](https://obamawhitehouse.archives.gov/the-press-office/2016/10/14/progress-african-american-commu
26	0	[1, 1, 1, 6]	[1, 1, 1, 6,	During his 8 years in office	Though he created some new programs that may have enfranchised people of colour, Obama also allowed the continuation of established ones that had
27	1	[1, 1, 1, 6]	[1, 4, 4]	During his 8 years in office	The Obama administration has repeatedly failed in the Middle East.
28	0	[1, 1, 1, 6,	[1, 1, 1, 6,	Though he created some n	In 2013, the Obama administration [freed up money](https://www.rollcall.com/2013/05/15/job-corps-hits-turning-point-under-congress-watch/) from o
29	1	[1, 1, 1, 6,	[1, 1, 1, 6,	In 2013, the Obama admin	Though he created some new programs that may have enfranchised people of colour, Obama also allowed the continuation of established ones that had
30	0	[1, 1, 1, 6	[1, 1, 1, 6	In 2013, the Obama admin	The inspector general's 2018 report on Job Corps found that the training offered [no discernible long-term benefit](https://www.nytimes.com/2018/08/2

Sheet1

3.4 Data Cleaning:

The dataset was performing badly before cleaning; thus, I made some text preprocessing to reduce the noise during training. The text preprocessing includes make text lowercase, remove text in square brackets, remove links, remove punctuation and removing words containing numbers.

4 System description.

In my project, the goal is to train a neural-based NLP system that handles the Next Argument Prediction task on the dataset generated from the Kialo website. Instead of two single sentences the dataset is consist of two arguments. The system must be able to predict whether the second argument entails the first.

At the beginning I decided to solve the problem using the different models of Bert, eventually I extended the model with LSTM/GRU encoder-decoder model for experiment. I tried different pre-trained models, and I compared them using a baseline architecture which has pretrained Bert model and a simple linear classifier and later extended it. The metrics used to compare the models were Accuracy and F1.

The framework used was PyTorch, and Huggingface's Transformer library was exploited to use pre-trained models, such as BERT, ALBERT and DISTILBERT. The library also contains for each pre-trained model its tokenizer vocabulary which I used to produce inputs to the models.

4.1 Various components of the system:

4.1.1 Dataset splitting:

I downloaded around 300 files from kialo.com site and made ≈80,000 trainset from 190 files, ≈20,000 validation set from 70 files and 10,000 test set from 40 files. The files are stored in separate folders and the code goes through each folder to create trin, validation and test sets.

4.1.2 Training the Network:

- **Training**: This scrip performs several training epochs specified by the user (on the given training set) and, at the end of each training epoch, performs a validation process using the specified validation set. After training phase is ended the scrips also returns the loss and accuracy scores.
- **Computing answers**: Since I also want to simply compute answers for the sake of evaluating the model, I also wrote a script that uses a trained model to compute the answers to each sample of a dataset specified by the user. The output of the script is a text file containing a predicted answer for each input sample.
- **Evaluation**: At the end the evaluation script provides the performance score of each model using the F1 and Accuracy scores.

5. Architecture.

I tried three base architectures all of which are mainly based on pre-trained Transformers. All the following experiments are based on solving the Next Argument Prediction task as a classification problem. The first experiment is fine-tuning Bert that is linked to a linear binary classification network. The base model is based on a pre-trained language model (more specifically, BERT, DistilBERT and ALBERT). In the following phases I modified this baseline model, by adding or modifying the architecture on top of the base model (i.e., LSTM and GRU encoder-decoder).

5.1 Hyper parameters:

These models were trained using the same hyperparameters and using the same train, validation, and test dataset to maintain coherence between the different experimentations that will be described in the next section. There is a configuration file to be used among all the experiments; the main parameters were the same (i.e., number of epochs, learning rate etc.). I tested each model by running a 5 training epoch, using a learning rate of 2e-5 and a dropout rate of 0.2. I used Cross Entropy loss and AdamW optimizer.

5.2 Experiment Phase 1

Base Model as version 0 In experiment phase 1 all the base models are trained and validated with an extension of a single Linear classification layer. In this case the parameters of the BERT models are frozen [there is no correction of Bert weights during back propagation] except for the linear classification layer.

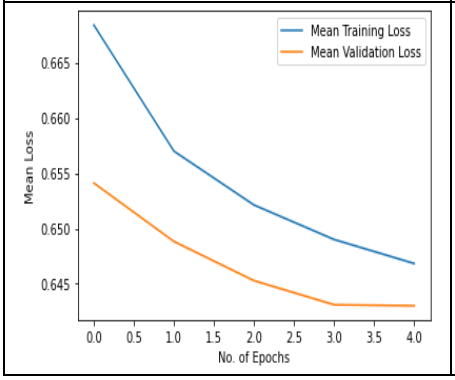
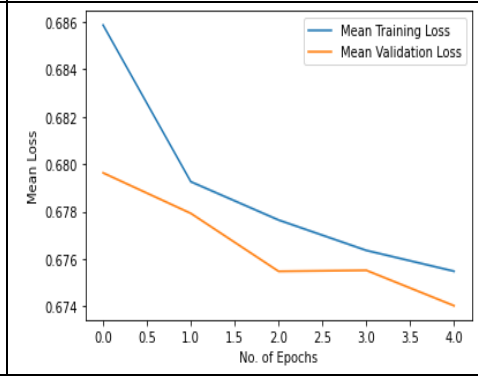
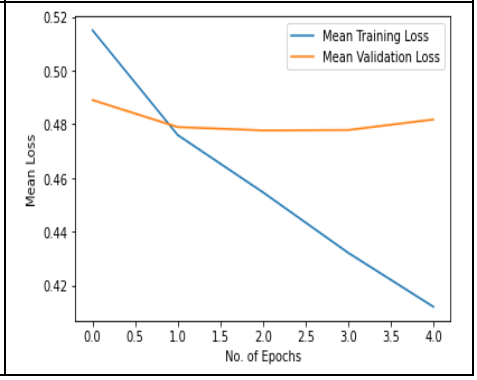
- **BERT Base Model v0**: I created a baseline model consisting of a BERT pre-trained layer with a simple classification layer on top of it which was taught to predict the answer whether the next argument is entailment or not.

- **DistilBERT Base model v0:** the architecture seen before is used with DistilBERT as pre-trained model. DistilBERT is a smaller version of BERT, thanks to a smaller number of used Transformers (6 compared to 12 of the base BERT), so it is lighter and faster to train.
- **ALBERT Base model v0:** the same architecture seen before is used with ALBERT as pre-trained model. ALBERT has a great achievement in terms of memory and computational cost in many tasks by Decoupling the embedding dimension from the dimension of the hidden layers to avoid an embedding matrix.

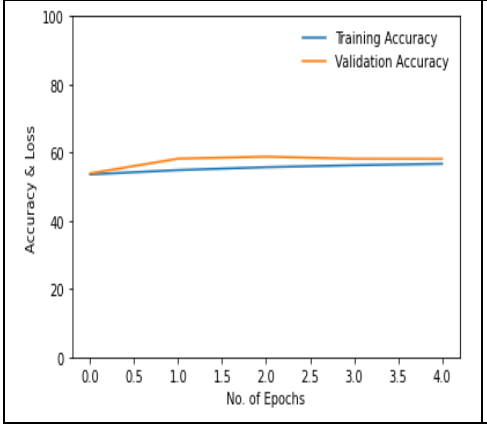
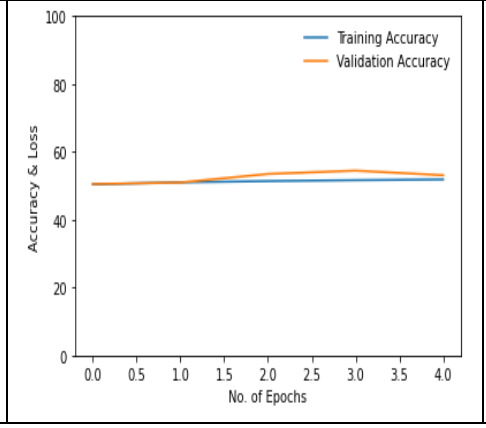
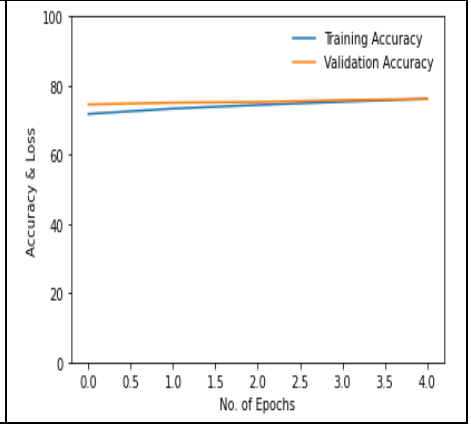
5.3 Experiment Phase 2

Modified Base Model v1 [LSTM]: In experiment phase 2 I added to all the base models an LSTM encoder-decoder architecture before the linear classifier. Though the Network got little deeper than the naïve feed forward network on top of the Bert embedding, the accuracy improvement was not as much significant.

Mean Loss:

Bert-base + LSTM	*DistilBert+LSTM*	*Albert+LSTM*
		

Accuracy:

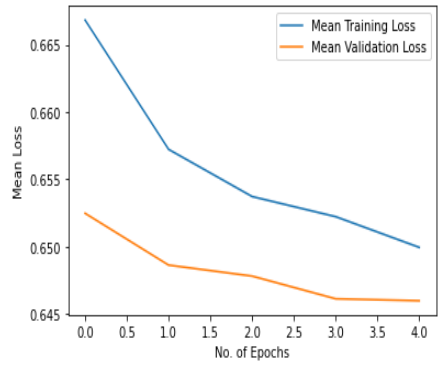
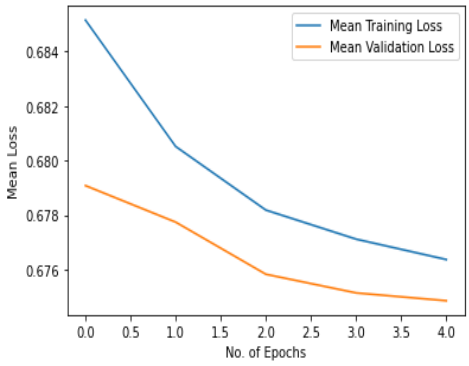
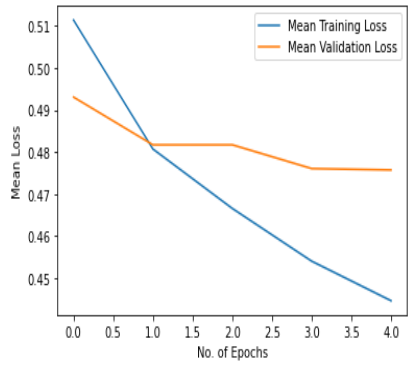
Bert-base + LSTM	*DistilBert+LSTM*	*Albert+LSTM*
		

5.4 Experiment Phase 3

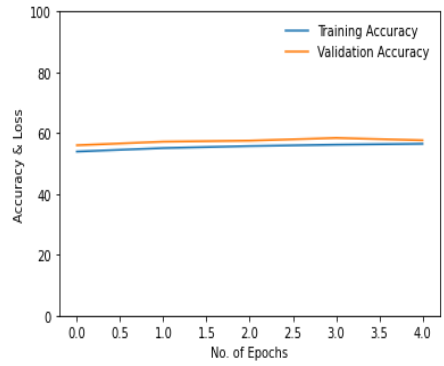
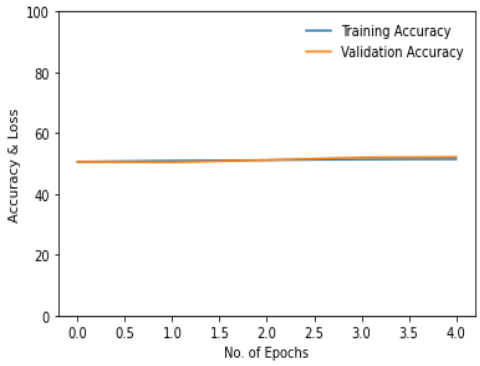
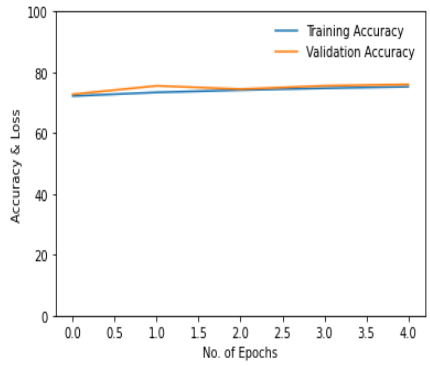
Modified Base Model v2 [GRU]: In experiment phase 3 I added to all the base models an GRU encoder-decoder architecture before the linear classifier and Bert embedding. The network got better relative to the previous networks in execution time and accuracy performance.

The following graphs are the output of the experiments:

Mean Loss:

Bert-base + GRU	*DistilBert+GRU*	*Albert+GRU*
		

Accuracy:

Bert-base + GRU	*DistilBert+GRU*	*Albert+GRU*
		

5.4 Conclusion I:

Even though the models were trained with the same parameter setting, they all behave somehow differently. BERT-base took more time to train, though the loss is not promising. DistilBert was much faster than BERT-base, but the loss and accuracy was worse than BERT-base. ALBERT in the other hand took less time than BERT-base and the loss decreased, and accuracy increased significantly.

6. Test Phase.

6.1 Test Output:

I checked the performances of the various models on the same test set. From the 3 phases of experiment the first phase (base model version 0) and second phase version 2 is aborted due to a poor model performance, so it is not included in the following analysis. Better models are pre-trained models with GRU. The best performing model was ALBERT modified v2, with an Accuracy score of 92% and a F1 score of 92%.

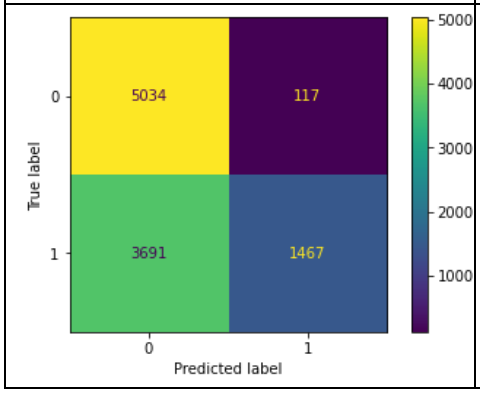
6.2 Confusion Matrix:

Looking at the confusion matrix I noticed the models were performing badly in classifying entailment of the two arguments when they are labeled as 1 [i.e. “Not next argument”]. I found out that the fact that I used the transpose of the arguments as label 1 ”NOT NEXT ARGUMENT”, is creating confusion during learning process. I came to this conclusion because when I run the model with test set that doesn' have a transposed sentence the performance increased significantly, and thus I trained an entire network without the transposed argument alignmnet during dataset creation.

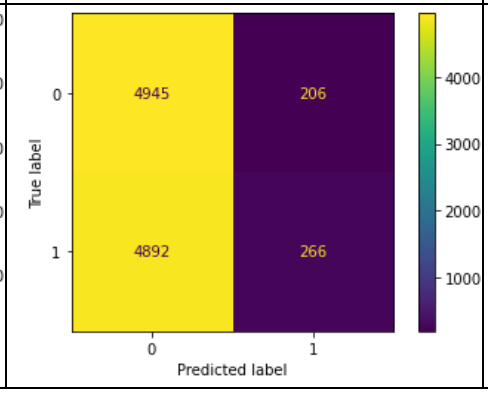
The following confusion matrix shows the difference between the models trained and tested with or without the datset that contains a transposed argument pairing.

With Transposed dataset:

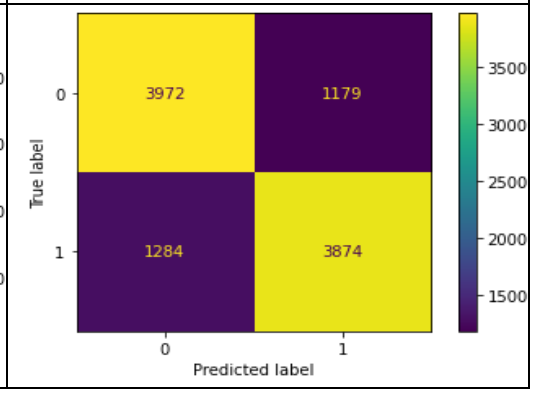
Bert-base + GRU



DistilBert+GRU

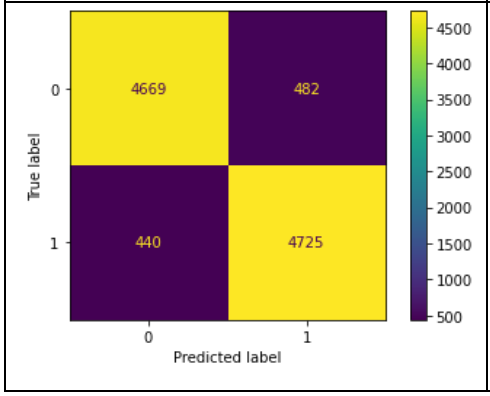


Albert+GRU

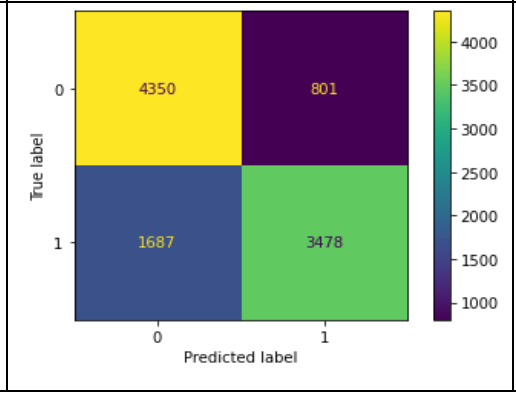


Without Transposed dataset:

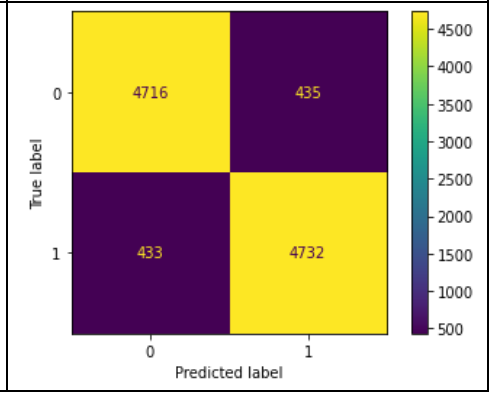
Bert-base + GRU



DistilBert+GRU



Albert+GRU



6.3 F1 and Accuracy Score:

The following table resumes the experiments scores.

Model	+ Transpose		- Transpose	
	Accuracy	F1	Accuracy	F1
Bert-base + GRU	63%	72%	91%	91%
DistilBert + GRU	51%	65%	78%	76%
Albert + GRU	76%	76%	92%	92%

6.2 Conclusion II:

All these experiments were done using the same training, validation, and testing sets except for the part I removed the transposed dataset in the last experiment. In the experiment DistilBert was performing the worst, but the improvement was quite impressive when I removed the transposed alignment of the arguments as not next argument. Just to test even more the model, I also created manually a very small test set by picking one file randomly from kialo.com and by feeding the arguments to the network. The Albert + GRU network was performing well and answered correctly to most of the classification.

6.3 Discussion:

- A major improvement would be to add a residual connection from the output of the pre-trained layer to the input of the LSTM/GRU decoder to help preserve the information after the LSTM/GRU encoder.
- Another test that could be done is to train these models for a higher number of epochs to see were these models reach their peak performances. Some hyperparameter tuning may also increase the scores further.

7. Bibliography References.

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” CoRR, vol. abs/1706.03762, 2017.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” CoRR, vol. abs/1810.04805, 2018.

[3] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,”

[5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” 2019.

[6] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “Transformers: State-of-the-art natural language processing,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, (Online), pp. 38–45, Association for Computational Linguistics, Oct. 2020.